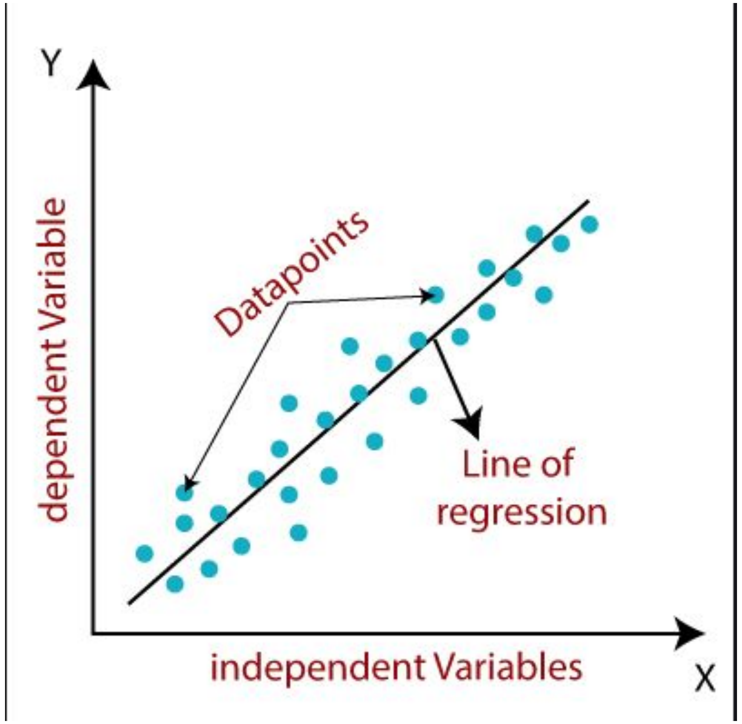
Regression

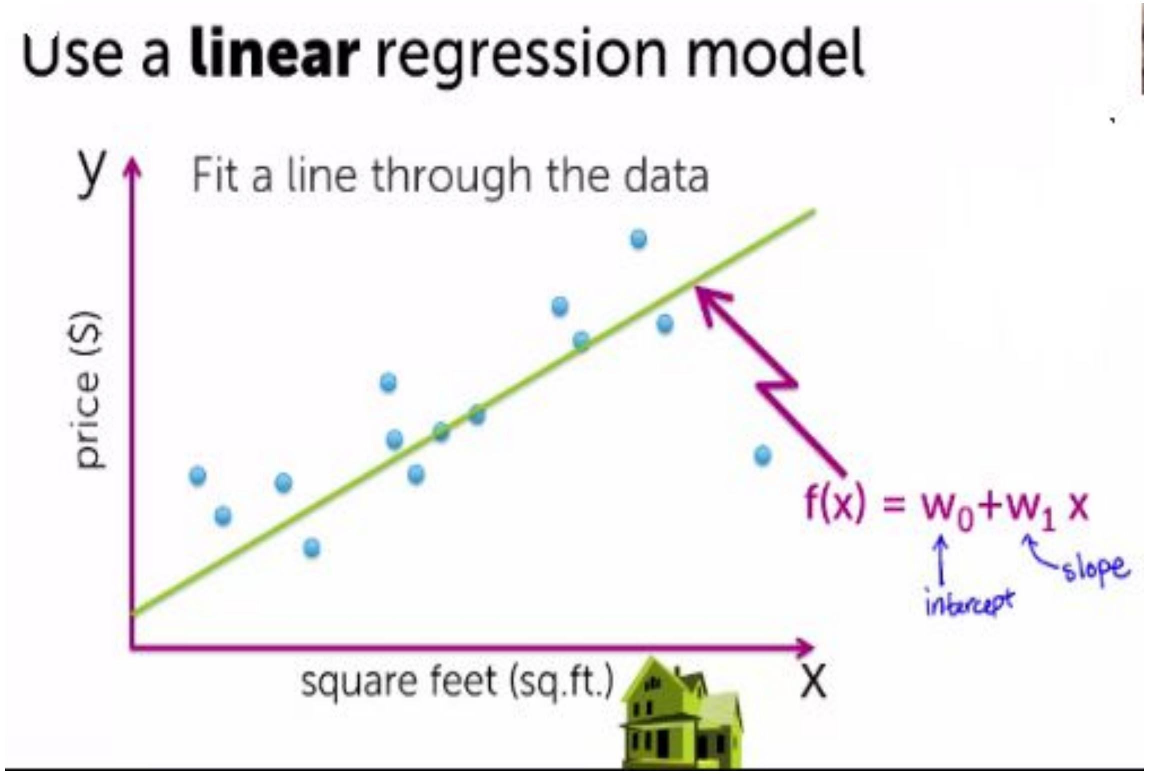
Regression

- A regression problem is the problem of determining a relation between one or more independent variables and an output variable which is a real continuous variable, given a set of observed values of the set of independent variables and the corresponding values of the output variable.



Examples

- Let us say we want to have a system that can predict the price of a used car.
- Inputs are the car attributes,brand, year, engine capacity, mileage, and other information that we believe affect a car's worth.
- The output is the price of the car.



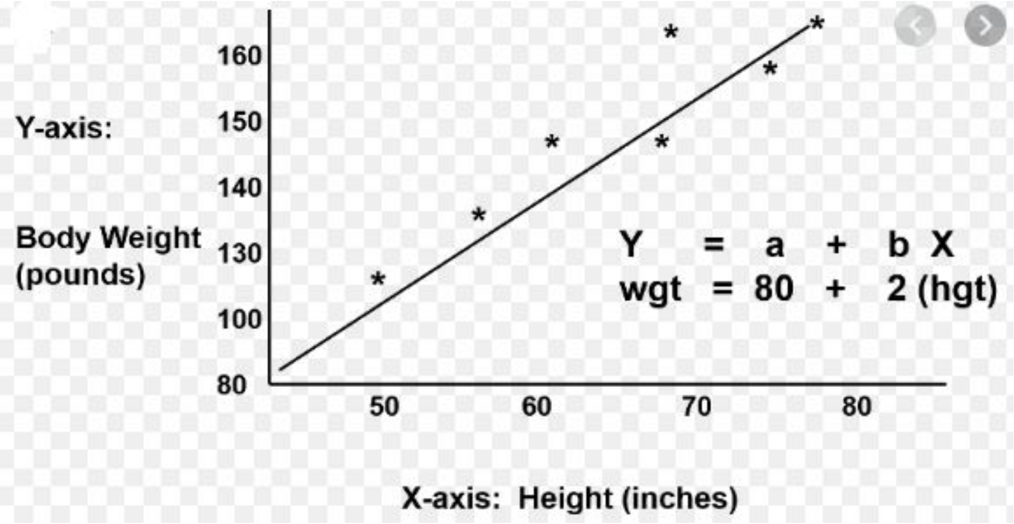
Different Regression Models

- The different regression models are defined based on type of functions used to represent the relation between the dependent variable y and the independent variables.

- Simple linear regression
- Multivariate linear regression
- Polynomial Regression
- Logistic Regression

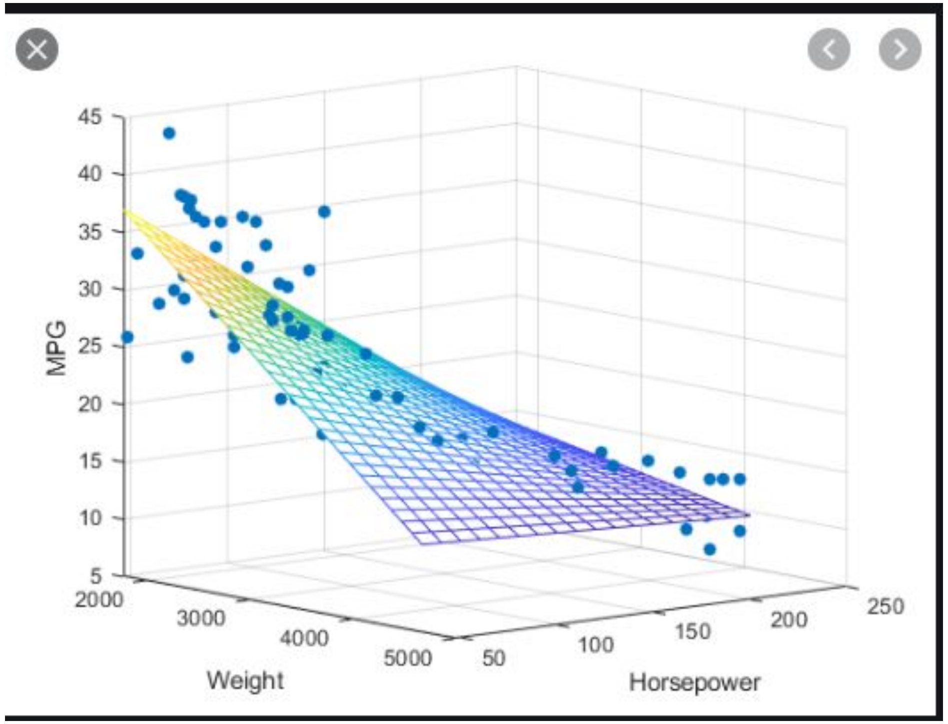
Simple linear regression

- 1. Simple linear regression
- Assume that there is only one independent variable x. If the relation between x and y is modeled by the relation
  - **y = a + bx**
- then we have a simple linear regression.



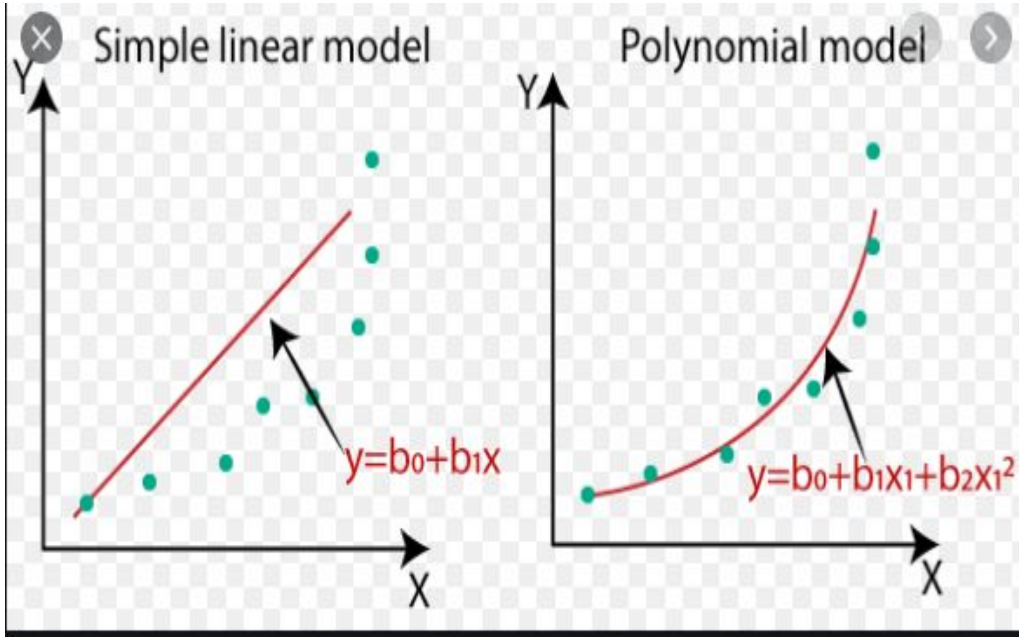
Multivariate linear regression

- There are more than one independent variable, say x1; : : : ; xn,
- and the assumed relation between the independent variables and the dependent variable is
- $y = a_0 + a_1x_1 + \ldots\ldots.. + a_nx_n$:
- .



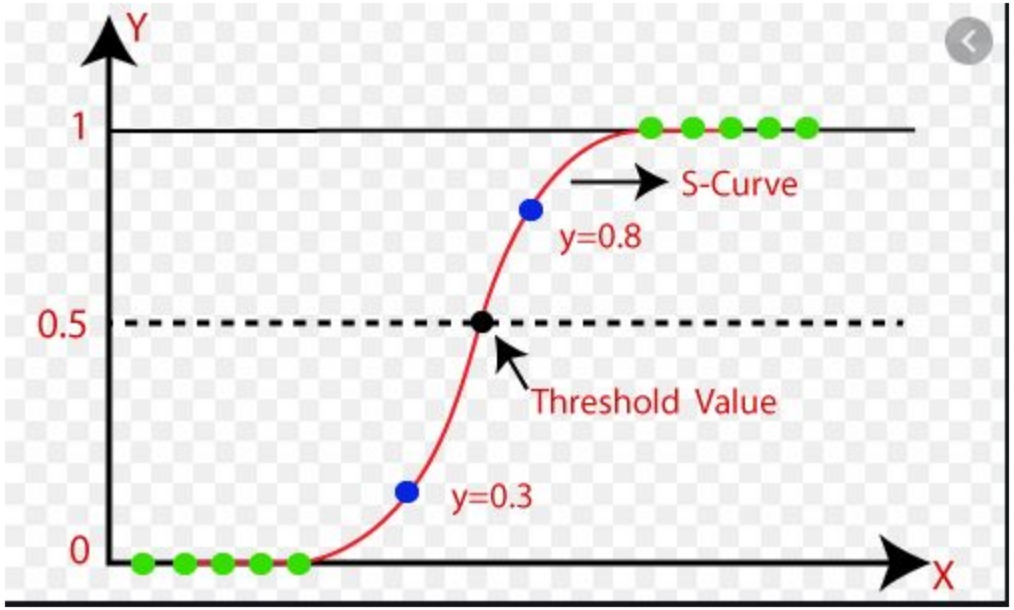
Polynomial Regression

- There is only one continuous independent variable x and the assumed model is
- $y = a_0 + a_1x + \ldots + a_nx^n$:



Logistic Regression

- Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature.
- Even though the output is a binary variable, what is being sought is a probability function which may take any value from 0 to 1.



Simple linear regression

- Let x be the independent predictor variable and y the dependent variable. Assume that we have a set of observed values of x and y:
- A simple linear regression model defines the relationship between x and y using a line defined by an equation in the following form:

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

- In order to determine the optimal estimates of α and , β in estimation method known as Ordinary Least Squares (OLS) is used.

- In the OLS method, the values of y-intercept and slope are chosen such that they minimize the sum of the squared errors; that is, the sum of the squares of the vertical distance between the predicted y-value and the actual y-value .
- Let ^ yi be the predicted value of yi. Then the sum of squares of errors is given by

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Table 7.1: Data set for simple linear regression

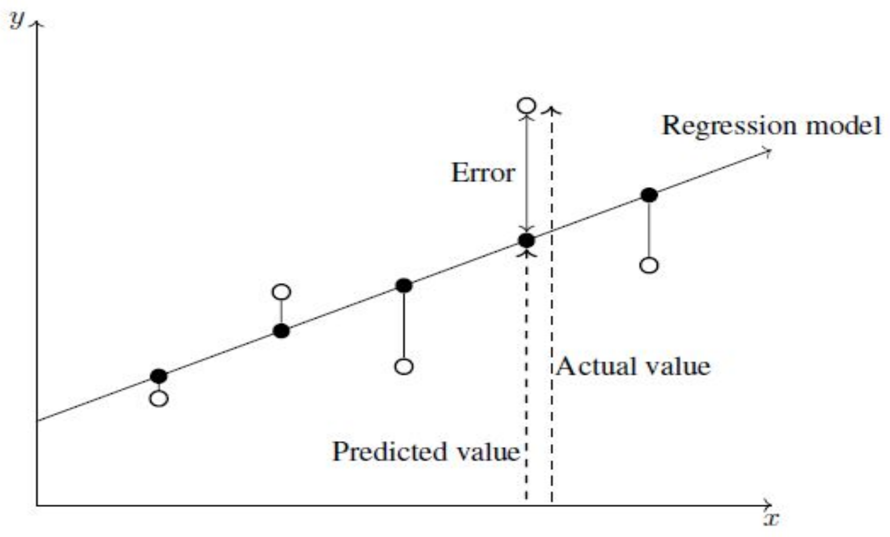


Figure 7.1: Errors in observed values

So we are required to find the values of α and β such that E is minimum. Using methods of calculus, we can show that the values of a and b, which are respectively the values of α and β for which E is minimum, can be obtained by solving the following equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

Formulas to find a and b

Recall that the means of x and y are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$
$$\bar{y} = \frac{1}{n}\sum y_i$$

and also that the variance of x is given by

$$\mathrm{Var}\,(x) = \frac{1}{n-1}\sum (x_i - \bar{x_i})^2.$$

The *covariance of x and y*, denoted by $\mathrm{Cov}(x, y)$ is defined as

$$\mathrm{Cov}(x, y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

It can be shown that the values of a and b can be computed using the following formulas:

$$b = \frac{\mathrm{Cov}\,(x, y)}{\mathrm{Var}\,(x)}$$
$$a = \bar{y} - b\bar{x}$$

Example

Obtain a linear regression for the data in Table 7.2 assuming that y is the independent variable.

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

Table 7.2: Example data for simple linear regression

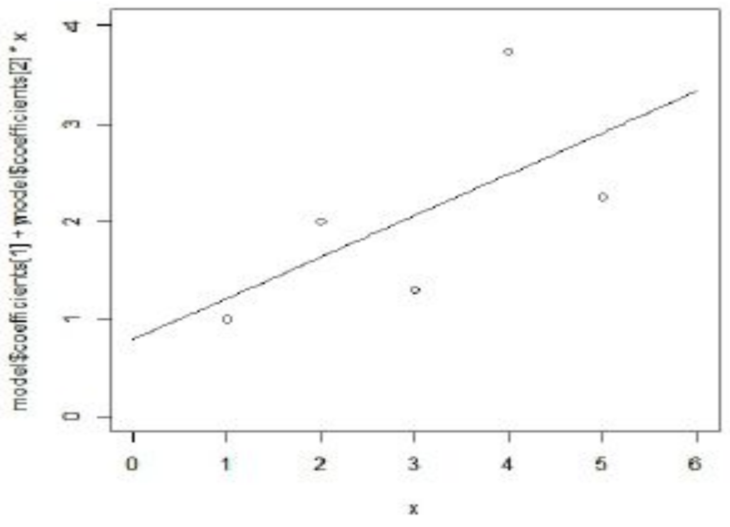


Figure 7.2: Regression model for Table 7.2

Solution

In the usual notations of simple linear regression, we have

$$
\begin{aligned}
n &= 5\\
\bar{x} &= \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)\\
&= 3.0\\
\bar{y} &= \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)\\
&= 2.06\\
\mathrm{Cov}\,(x, y) &= \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]\\
&= 1.0625\\
\mathrm{Var}\,(x) &= \frac{1}{4}[(1.0 - 3.0)^2 + \cdots + (5.0 - 3.0)^2]\\
&= 2.5\\
b &= \frac{1.0625}{2.5}\\
&= 0.425\\
a &= 2.06 - 0.425 \times 3.0\\
&= 0.785
\end{aligned}
$$

Therefore, the linear regression model for the data is

$$y = 0.785 + 0.425x.$$

Question-1

a) The following table shows the midterm and final exam grades obtained for students in a database course. (6)

X Midterm exam	Y Final exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(i) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

(ii) Predict the final exam grade of a student who received an 86 on the midterm exam.

a) The following table shows the midterm and final exam grades obtained for students in a database course. (6)

X Midterm exam	Y Final exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(i) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

(ii) Predict the final exam grade of a student who received an 86 on the midterm exam.

(i) Step by step procedure $|D| = 12$; $\bar{x} = 866/12 = 72.167$; $\bar{y} = 888/12 = 74$. w1 = 0.5816 and w0 = 32.028 (4)

Solution y = 32.028 + 0.5816x. (1)

(ii) Predicted value y = 32.028 + (0.5816)(86) = 82.045 (1)