Linear regression

Part 1: Introduction and intuition of linear regression

By: Noureddin Sadawi, PhD

University of London

Correlation

- Quantifies the strength of a relationship between two numeric variables.
- Usually we are interested in the linear correlation between two numeric variables.
- ‘Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y.
- If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Correlation coefficient

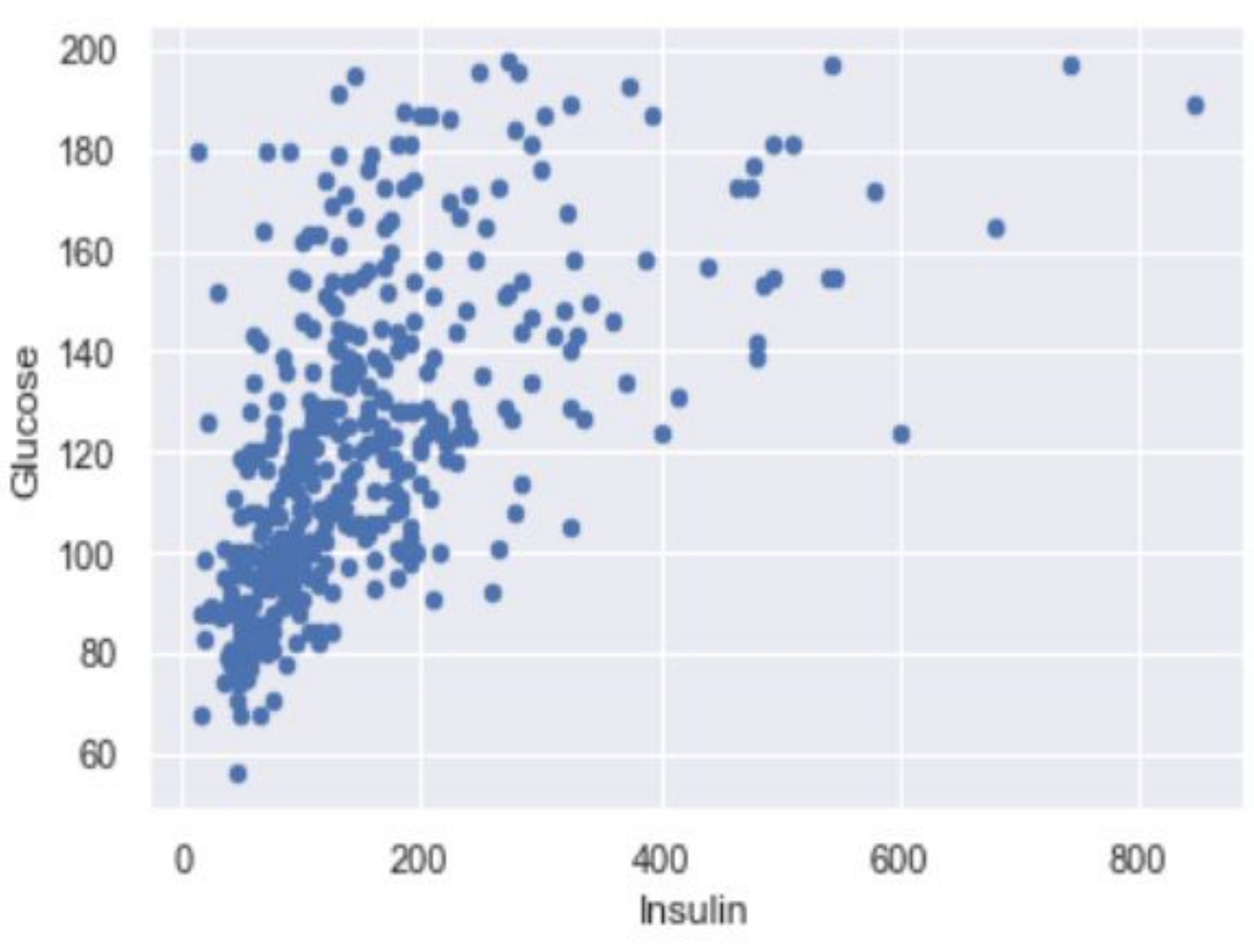
- ‘The correlation coefficient always lies between +1 (perfect positive correlation) and –1 (perfect negative correlation); 0 indicates no correlation.
- Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

$$r = \frac{Covar(x, y)}{\sqrt{Var(x)Var(y)}}$$

Correlation coefficient is not enough

- The correlation coefficient, r, quantifies the linear relationship between two variables.
- In other words, it tells us how close the points lie to a straight line.
- But it does not give us which straight line.



Introduction

- Often in Mathematics we are interested in variables that are deterministically related to one another.
- An example of such relationship is the one between degrees Fahrenheit and degrees Celsius which is given by the equation:

$$F = \frac{9}{5}C + 32$$

- Where F represents temperature in degrees Fahrenheit and C, temperature in degrees Celsius.
- If we know the temperature in degrees Celsius we can use this equation to compute it in degrees Fahrenheit.

Definition: linear regression

- Linear Regression involves using a linear approach to model the relationship between one dependent variable (or response variable) and one or more independent variables (or predictor variables) in order to gain information about one of them by knowing the values of the other(s).
- The independent variable is numeric.
- There are two types of linear regression: Simple linear regression and Multiple linear regression.

Simple and multiple linear regression

- Simple linear regression involves one dependent variable and one independent variable.
- Multiple linear regression is one involving one dependent variable and two or more independent variables.
- Regression can be used for prediction, estimation, modelling causal relationships and hypothesis testing.

The simple linear regression model

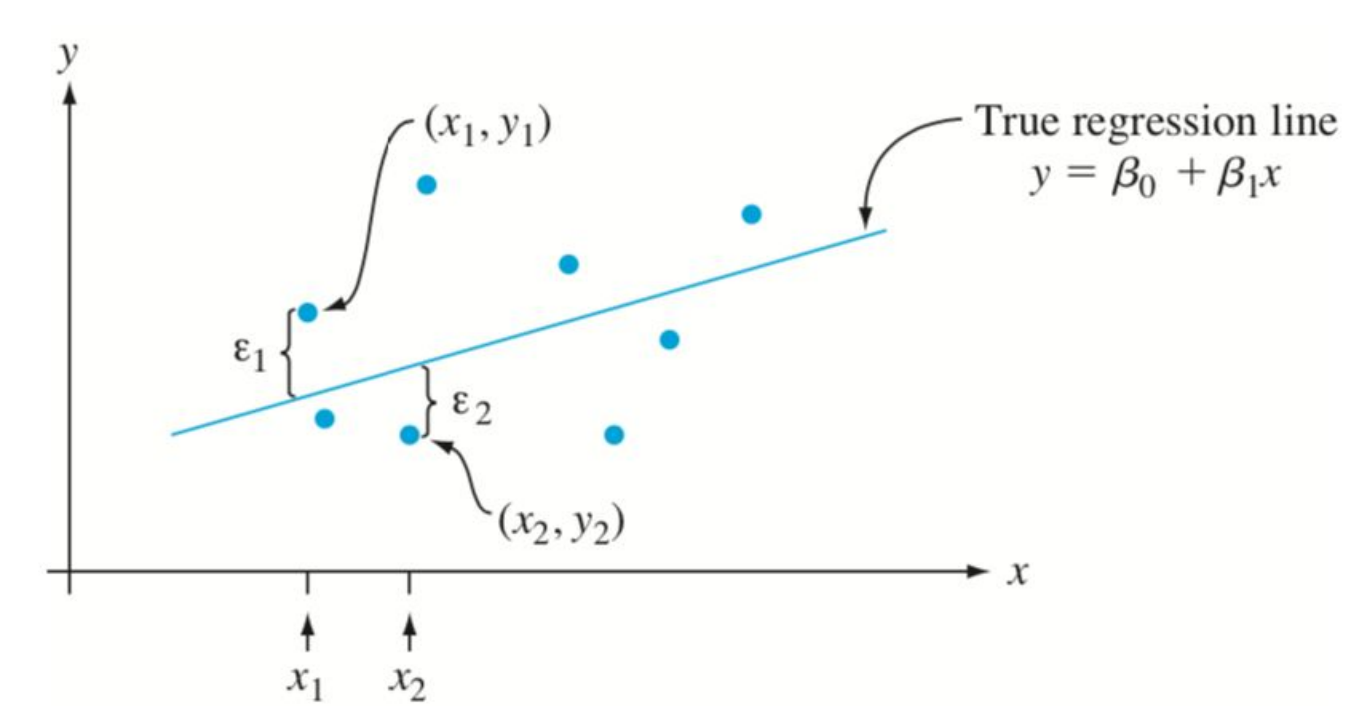
- The regression model for simple linear regression is given by:

$$Y = A + Bx + \epsilon$$

Where *A* and *B* are called the population parameters of the model, and ϵ is a random variable referred to as the error term.

- The error term ϵ accounts for the variability in *Y* that cannot be explained by the linear relationship between *X* and *Y*.
- Each distribution of *Y* values has its own mean or expected value.

The simple linear regression model



The equation that describes how the expected value of Y, denoted E(Y) or equivalently, E(Y|X=x), is related to X is called the *regression equation*.

The regression equation

- The regression equation for simple linear regression is given by:

$$E(Y) = A + Bx$$

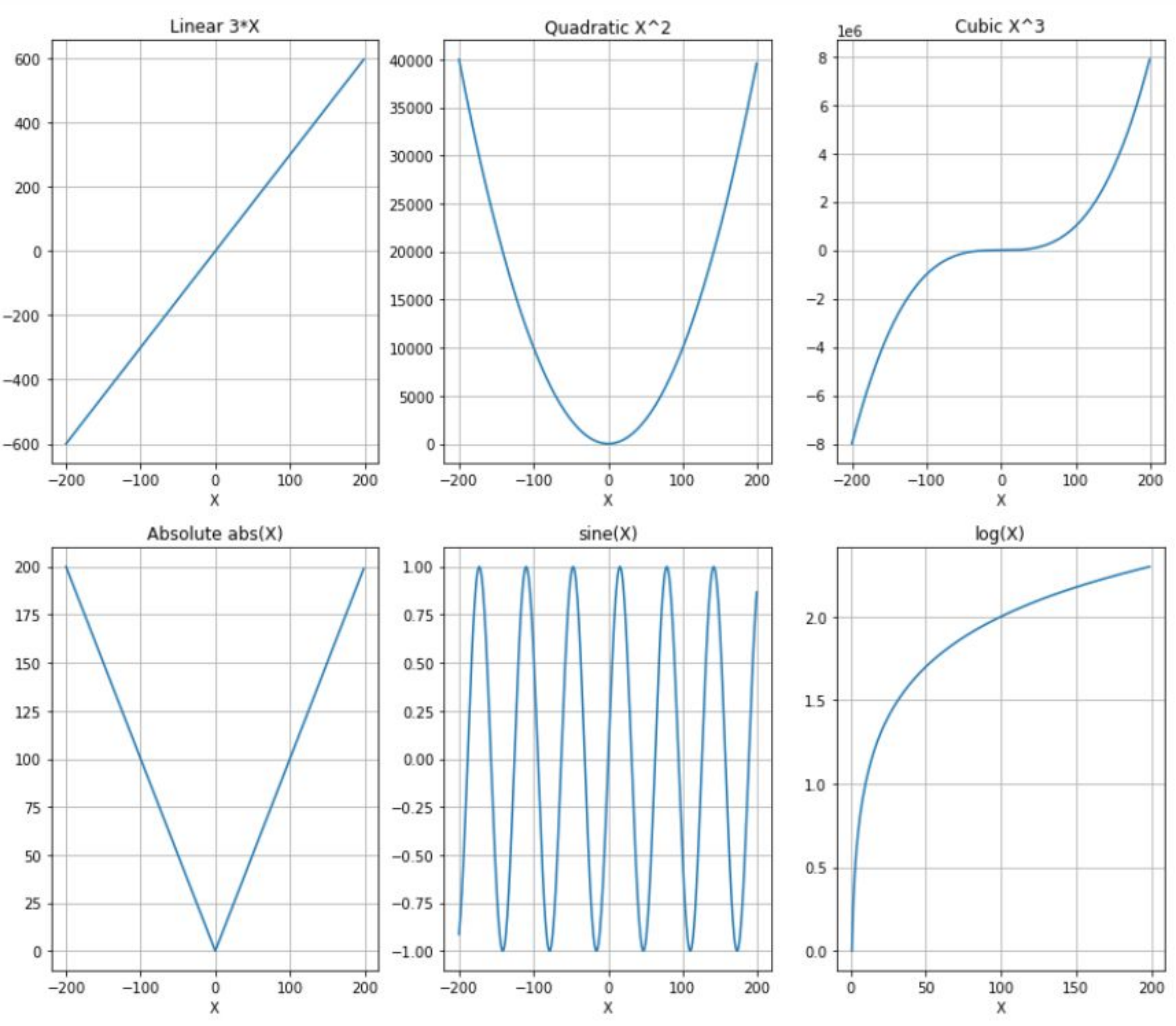
The graph of the simple linear regression equation is a straight line.

- *A* is the *y* intercept of the regression line.
- *B* is he slope or gradient.
- *E(Y)* is the mean or expected value of *Y* for a given value of *X*.

Example linear regression problems

- Remember in linear regression the outcome is usually a quantity (i.e. numeric).

- Analyse the relationship between the size, or area, of a house and its price (or predict the price from size or area).
- Predict Glucose level from Insulin level.
- Predict future company sales from available previous sales data.


Linear 3*X
600
400
200
0
−200
−400
−600
−200
−100
0
100
200
X
Quadratic X^2
40000
35000
30000
25000
20000
15000
10000
5000
0
−200
−100
0
100
200
X
Cubic X^3
1e6
8
6
4
2
0
−2
−4
−6
−8
−200
−100
0
100
200
X
Absolute abs(X)
200
175
150
125
100
75
50
25
0
−200
−100
0
100
200
X
sine(X)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−200
−100
0
100
200
X
log(X)
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
X