Logistic regression

Part 1: Overview of regression

By: Noureddin Sadawi, PhD

University of London

Regression

- We use regression to:
  - model the relationships between variables
  - allow predictions about future events.
- In regression the outcome is a quantity (i.e. a real number).
- Input variables can be real valued or discrete.
- When there are multiple input variables the problem is often denoted as a multivariate regression problem.
- If data observations are ordered by time the problem is called a time series forecasting problem.

Regression examples

- Demand analysis where we would like to predict the number of items a customer might buy.
- What is the association between school quality and academic achievement?
- The extent of association between smoking and heart disease status.
- The relationship between the monthly online sales and the online advertising costs.
- The relationship between the age and price for used cars.

Regression variables

There are two types of variables in regression:

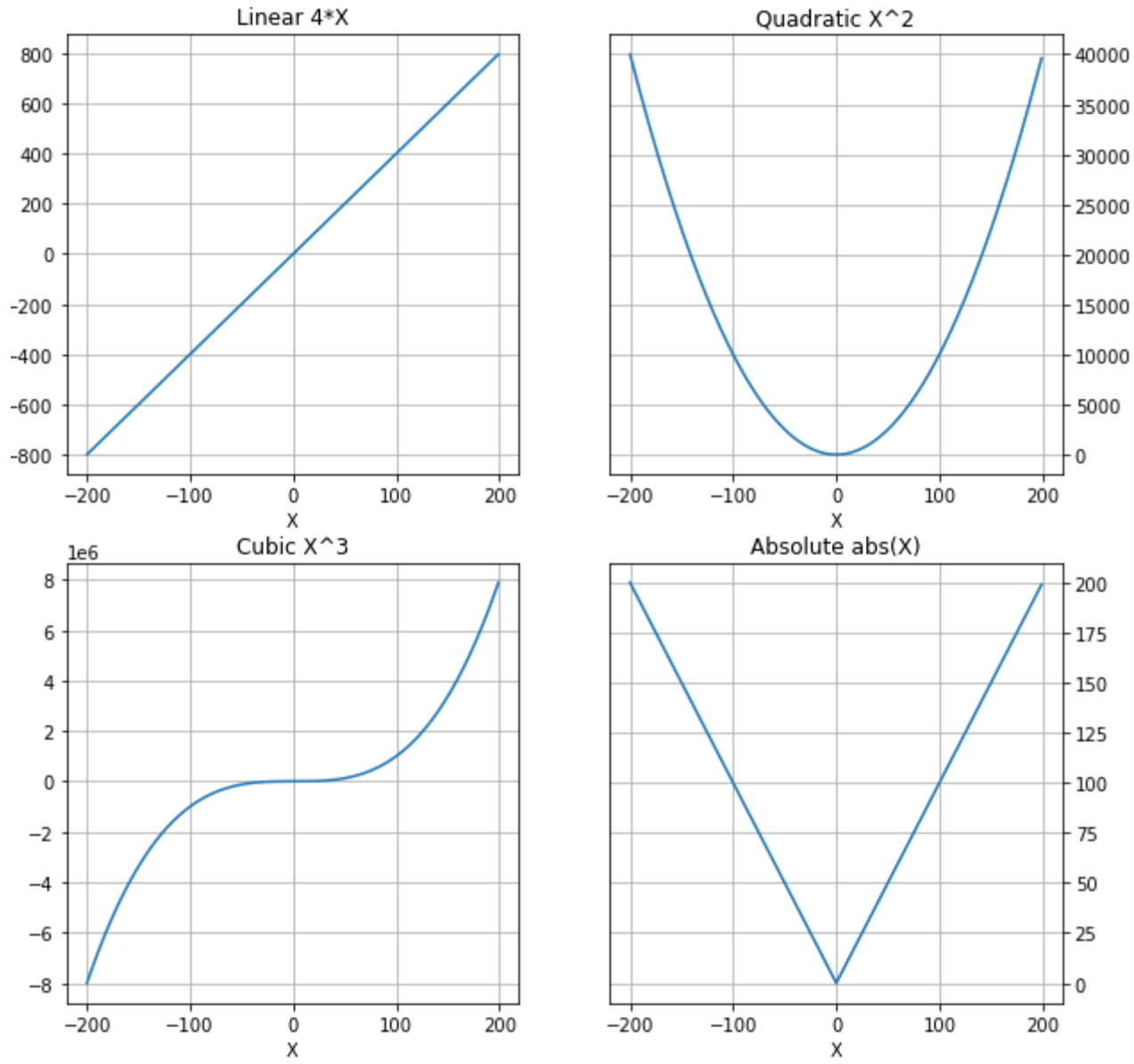
1. Response variable: the variable we want to investigate in relation to changes in other factors.
   - Sometimes referred to as the outcome, or the dependent variable.
2. Explanatory variables: variables that play a role in affecting the response variable.
   - Sometimes referred to as the independent variables, explanatory variables, covariates, predictors, attributes, features.

Regression models

The relationship between the independent variables and the dependent variable is expressed as a mathematical equation:

Dependent variable = function of (Independent variables)

$$Y = f(X)$$


Linear 4*X
800
600
400
200
0
-200
-400
-600
-800
−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

Regression models

- Coefficient estimates: size of effect of an independent variables on the outcome.
- P-value: the significance of that covariate.
- Confidence interval: if the procedure is repeated several times, this interval includes the true population coefficient value with a prescribed probability.