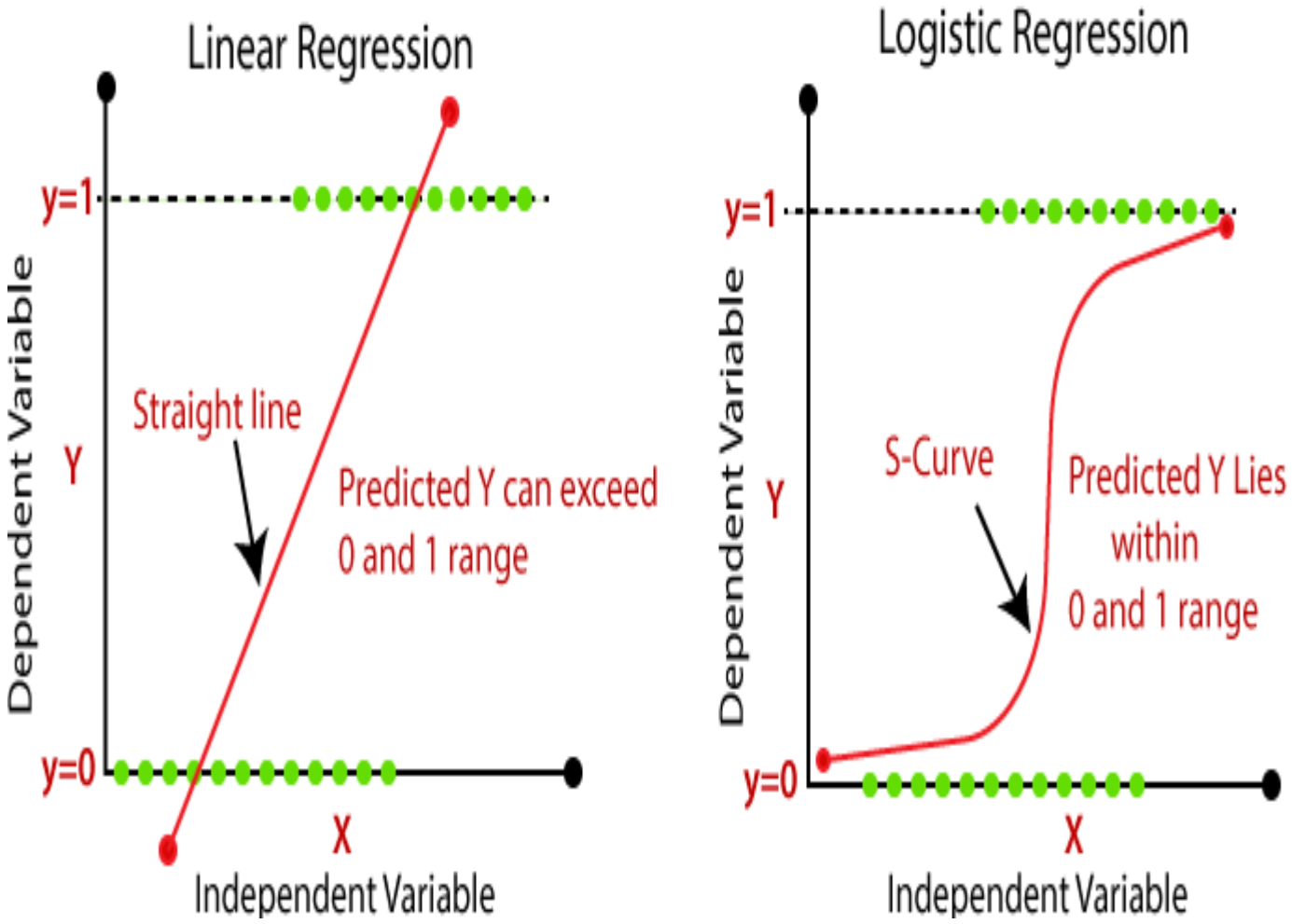
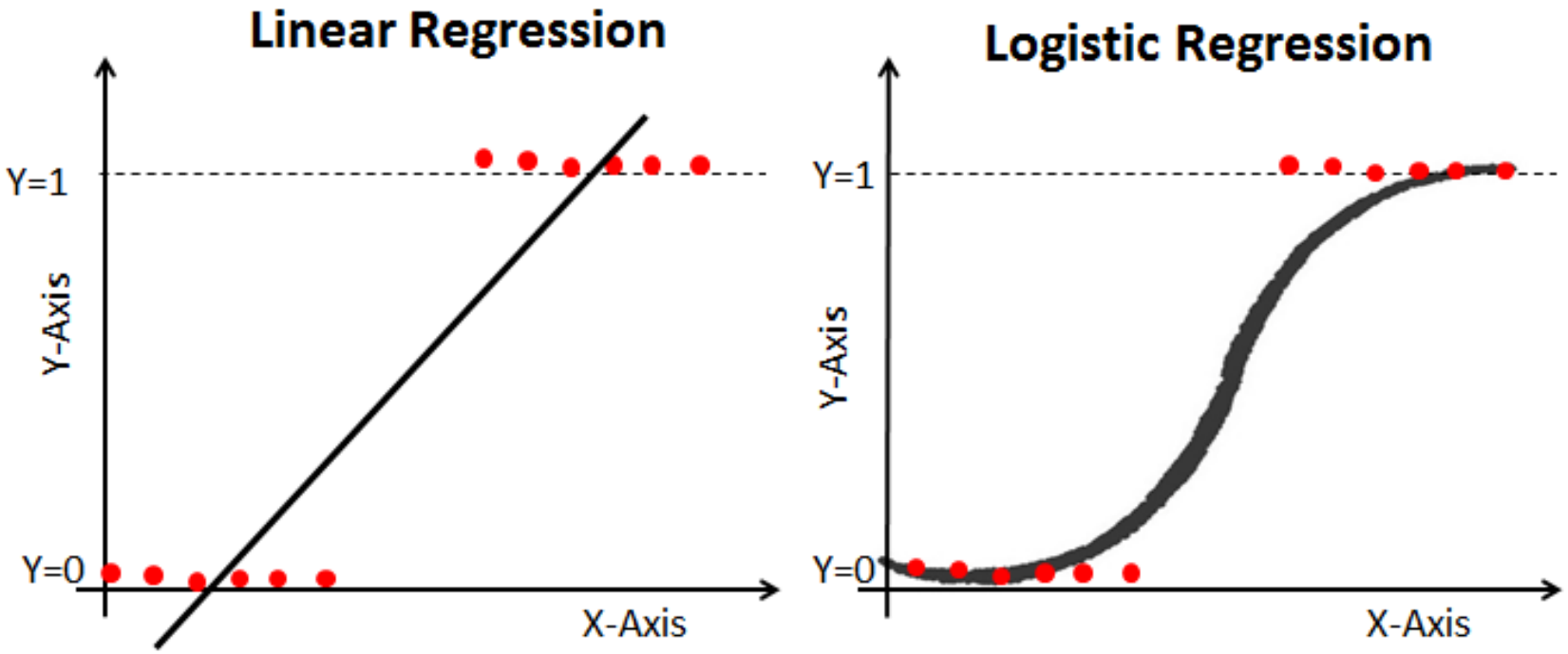
Linear versus Logistic Regression

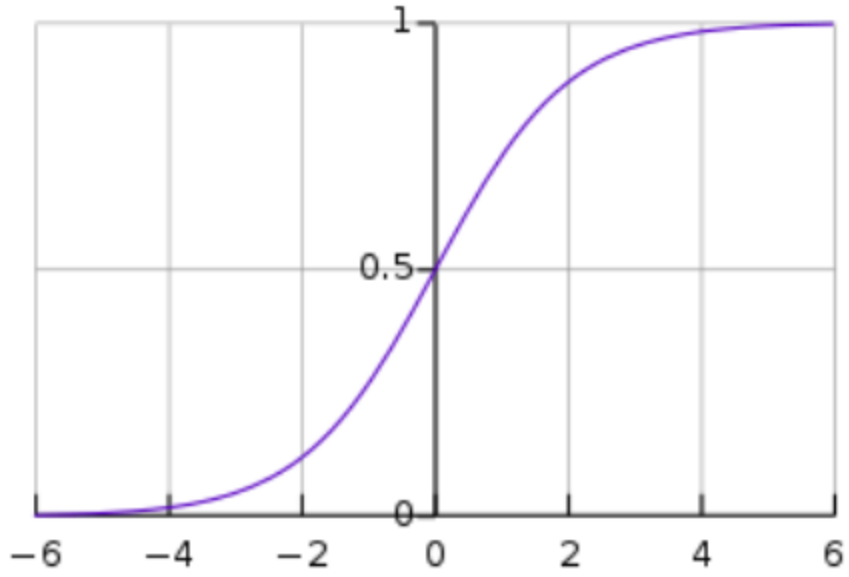
■Linear Regression	■Logistic Regression
■Target is an interval variable.	■Target is a discrete (binary or ordinal) variable.
■Input variables have any measurement level.	■Input variables have any measurement level.
■Predicted values are the mean of the target variable at the given values of the input variables.	■Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables.


Linear Regression
y=1
y=0
Dependent Variable
Y
Straight line
Predicted Y can exceed
0 and 1 range
X
Independent Variable
Logistic Regression
y=1
y=0
Dependent Variable
Y
S-Curve
Predicted Y Lies
within
0 and 1 range
X
Independent Variable

- Linear Regression – **can't take categorical variables** as input /independent variables. One needs to derive new fields.
- Linear regression **will have to ignore observations with missing values** of numeric independent variable
- Linear Regression – **will get hugely affected by outliers of numeric** independent variables
- For multiple independent variable case, linear regression will be little difficult to interpret

- Regression Tree – **can take categorical variables** as input / independent variables.
- Regression tree **can take missing as a class and can work easily** with the data to develop decision tree
- Regression Tree – **will not get affected by outliers of numeric** independent variables. It just makes class of numeric variable


Linear Regression
Y=1
Y-Axis
Y=0
X-Axis
Logistic Regression
Y=1
Y-Axis
Y=0
X-Axis



The equation for the sigmoid function is:

$$S(x) = \frac{1}{1 + e^{-x}}$$

It ensures that the generated number is always between 0 and 1 since the numerator is always smaller than the denominator by 1. See below:

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1}.$$