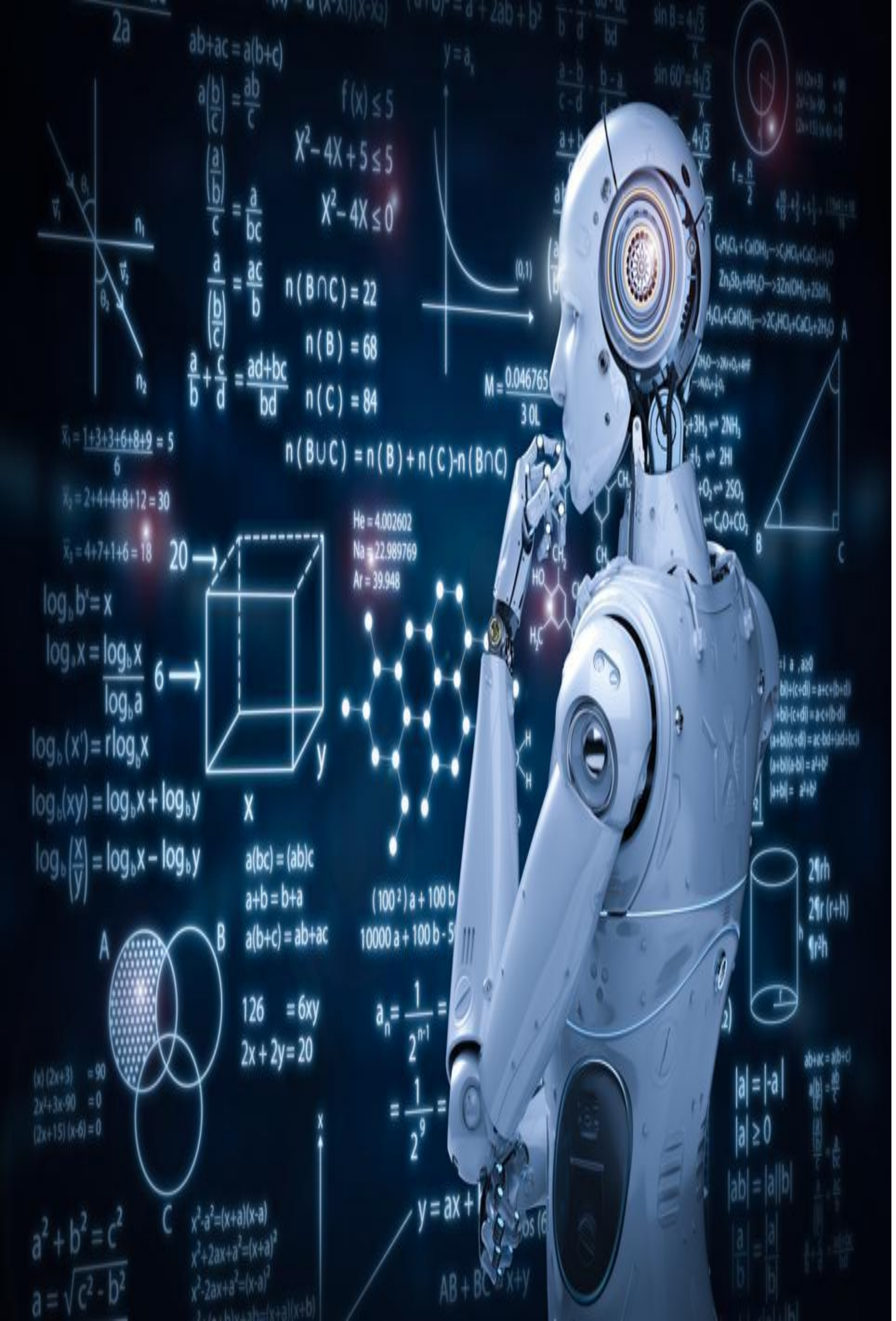


# Practical Machine Learning

## Day 4: SEP23 DBDA

Kiran Waghmare

# Agenda

- Regression
- Types of Regression

# Linear model

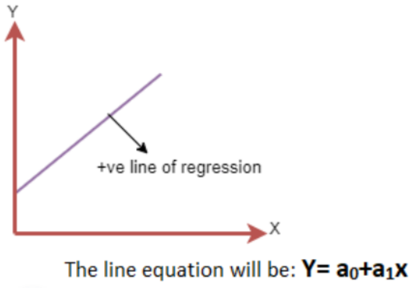
In regression, the relationship between Y and X is modelled in the following form:

$$Y = a + b * X + E$$

where:

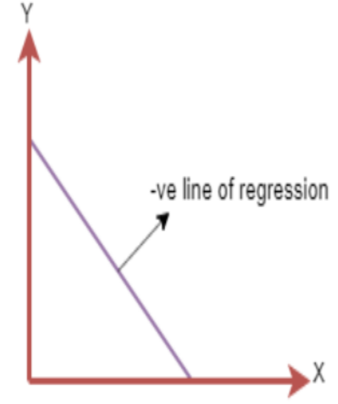
- **Y** is the dependent variable (Income in the example)
- **X** is the independent variable (IQ in the example)
- **a** is an intercept
- **b** is the coefficient
- **E** is an error term for each observation (since there is additional variation not explained by income)

# Linear Regression Line



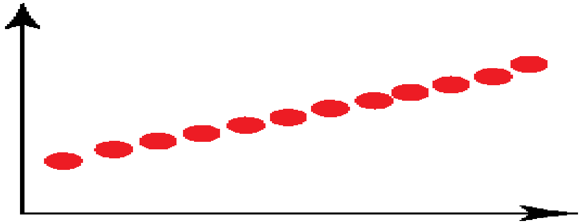

- A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

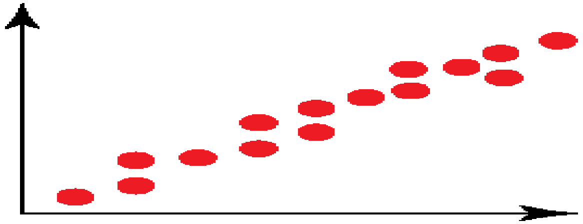



- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

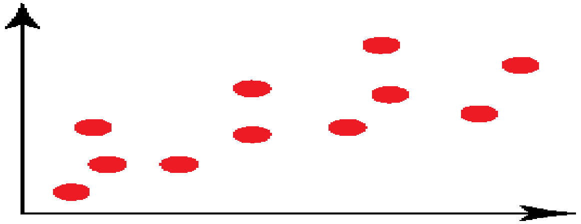
$$Y = \beta_0 + \beta_1 X + \varepsilon$$



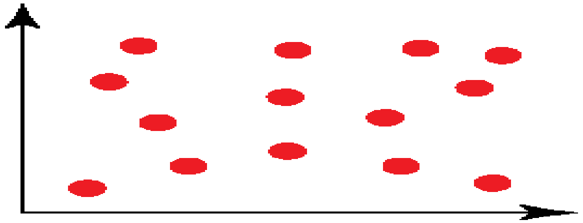
Perfect Positive Correlation



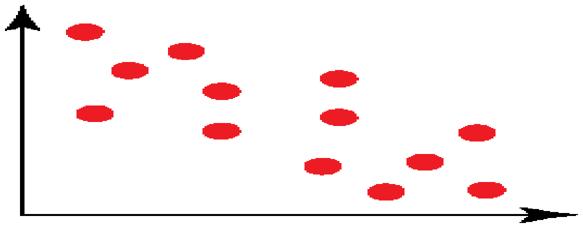
Strong Positive Correlation



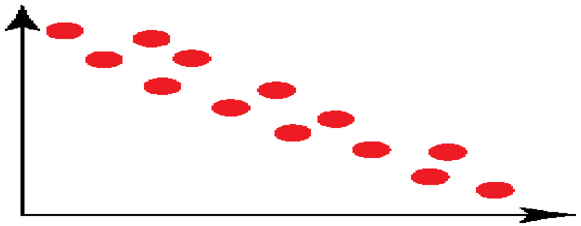
Weak Positive Correlation



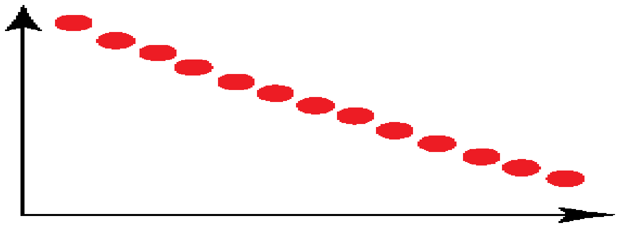
No Correlation



Weak Negative Correlation



Strong Negative Correlation



Perfect Negative Correlation

# 1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

## Residuals (regression error)

- Residuals or error in regression represents the distance of the observed data points from the predicted regression line

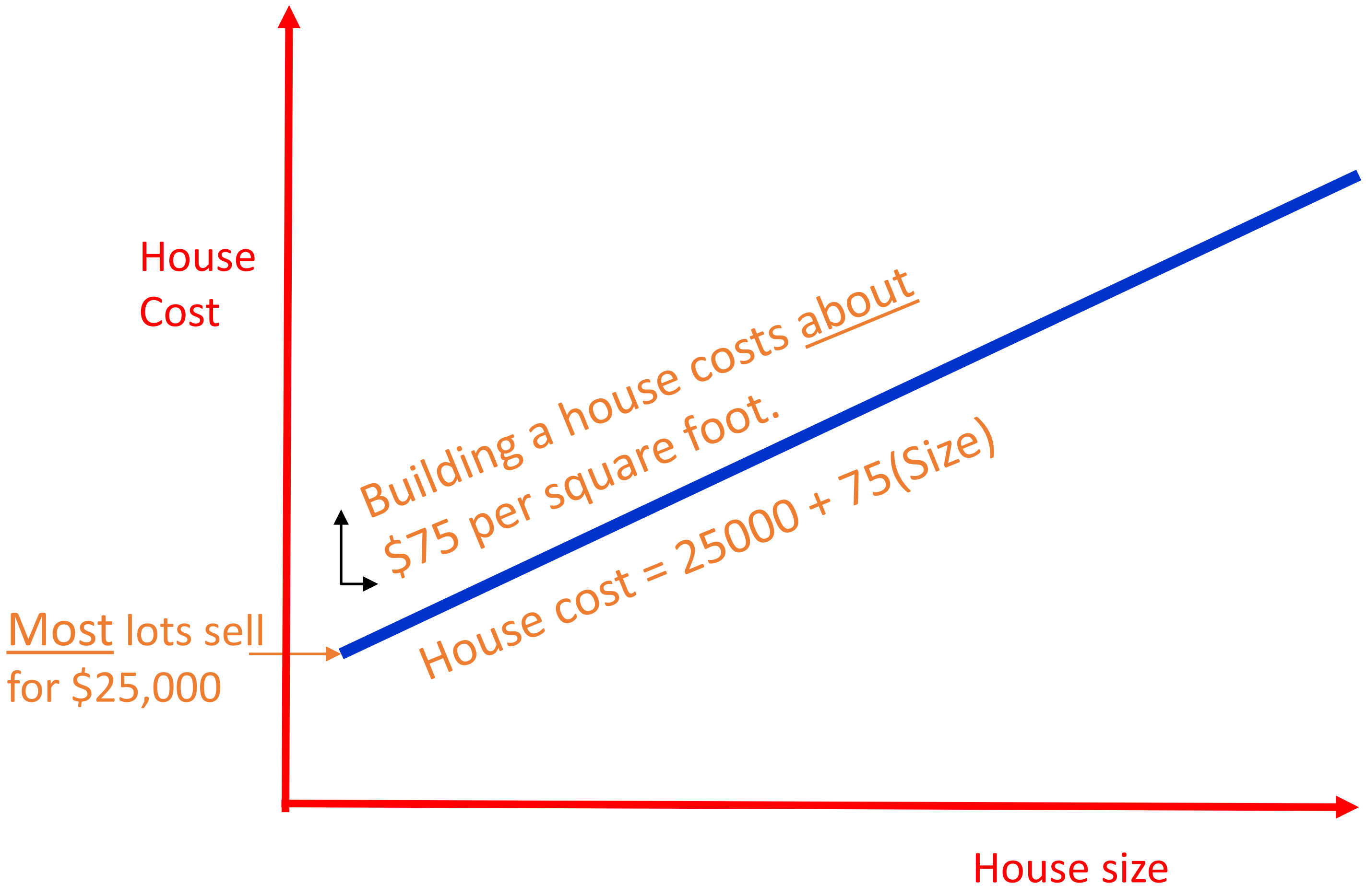
$$residuals = actual\ y(y_i) - predicted\ y\ (\hat{y}_i)$$

## Root Mean Square Error (RMSE)

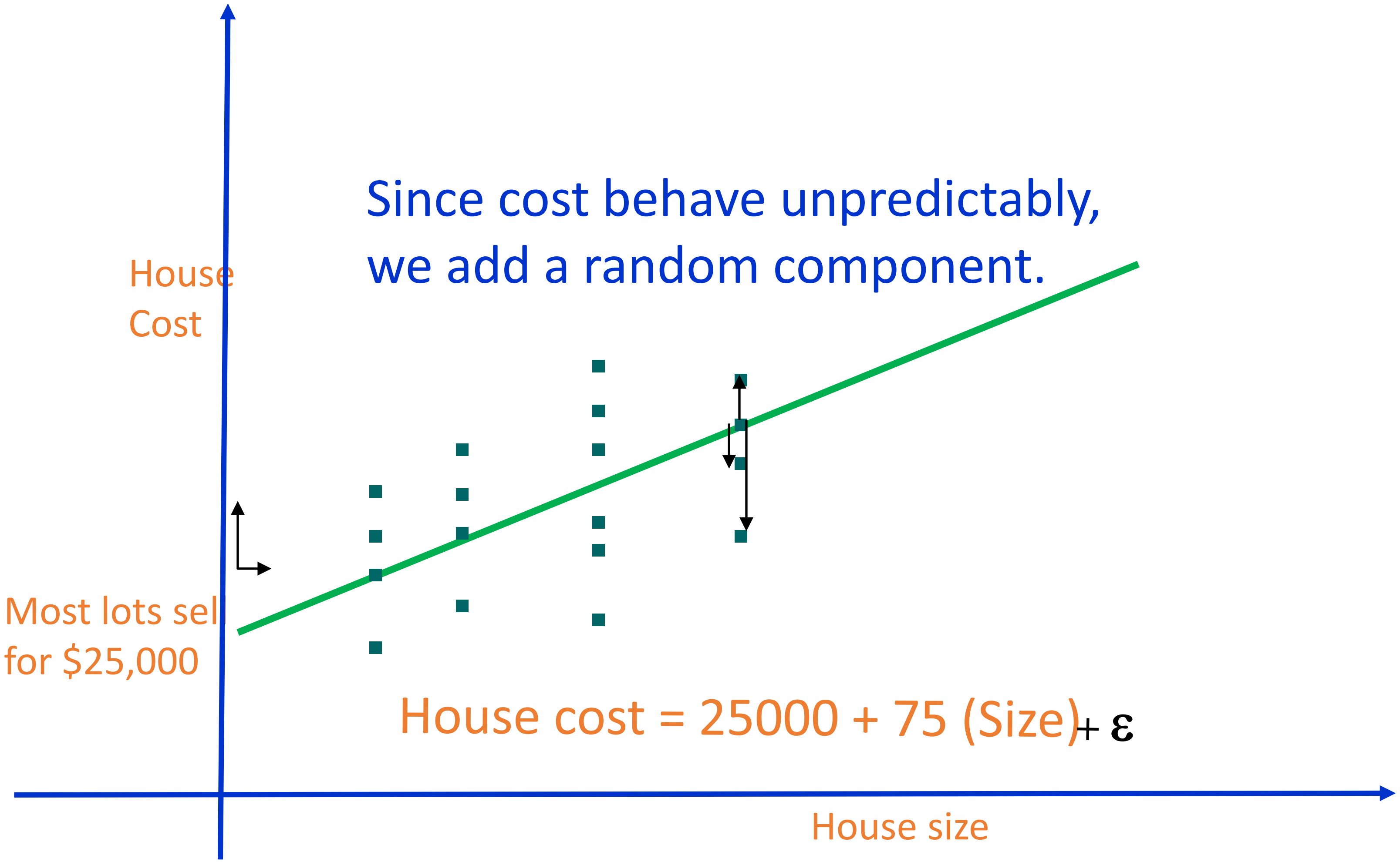
- RMSE represents the standard deviation of the residuals. It gives an estimate of the spread of observed data points across the predicted regression line.

# The Model

The model has a deterministic and a probabilistic components

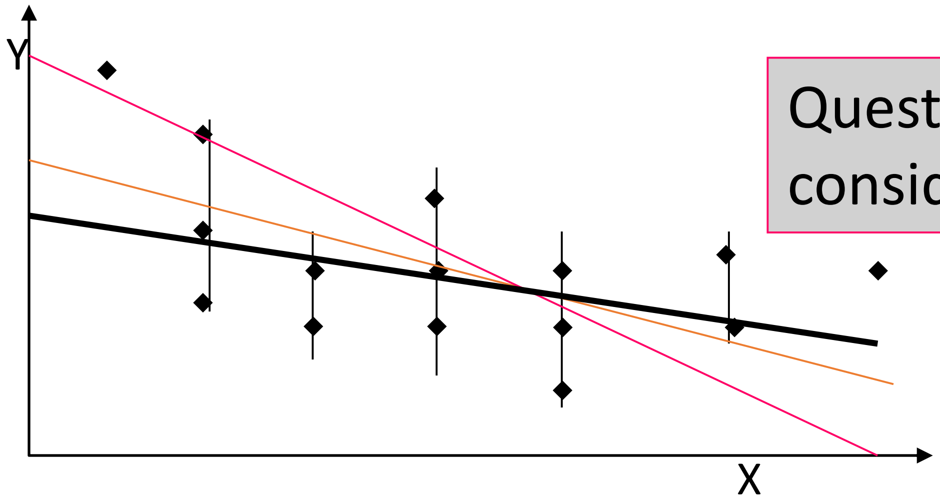


# However, house cost vary even among same size houses!



# Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.




Question: What should be considered a good line?

- $MeanSquaredError(mse) = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$
- $MeanAbsoluteError(mae) = (\frac{1}{n})\sum_{i=1}^{n}\left|y_i - x_i\right|$

# Gradient Descent:

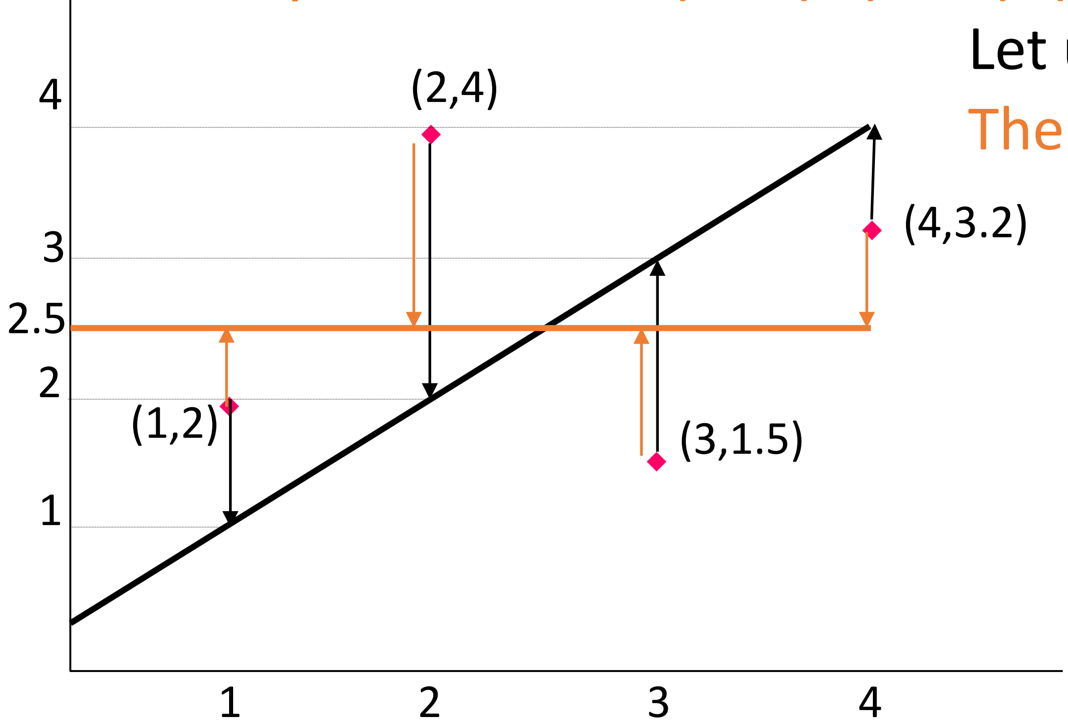
- Gradient descent is used **to minimize the MSE by calculating the gradient of the cost function.**
- A regression model uses gradient descent **to update the coefficients of the line** by reducing the cost function.
- It is **done by a random selection of values of coefficient** and then iteratively update the values to reach the minimum cost function.
- **Model Performance:**
- The Goodness of fit determines how the line of regression fits the set of observations.
- The process of **finding the best model out of various models** is called **optimization**.

$$\text{Sum of squared differences} = (2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$$

$$\text{Sum of squared differences} = (2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$$

Let us compare two lines

The second line is horizontal




The smaller the sum of squared differences the better the fit of the line to the data.

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

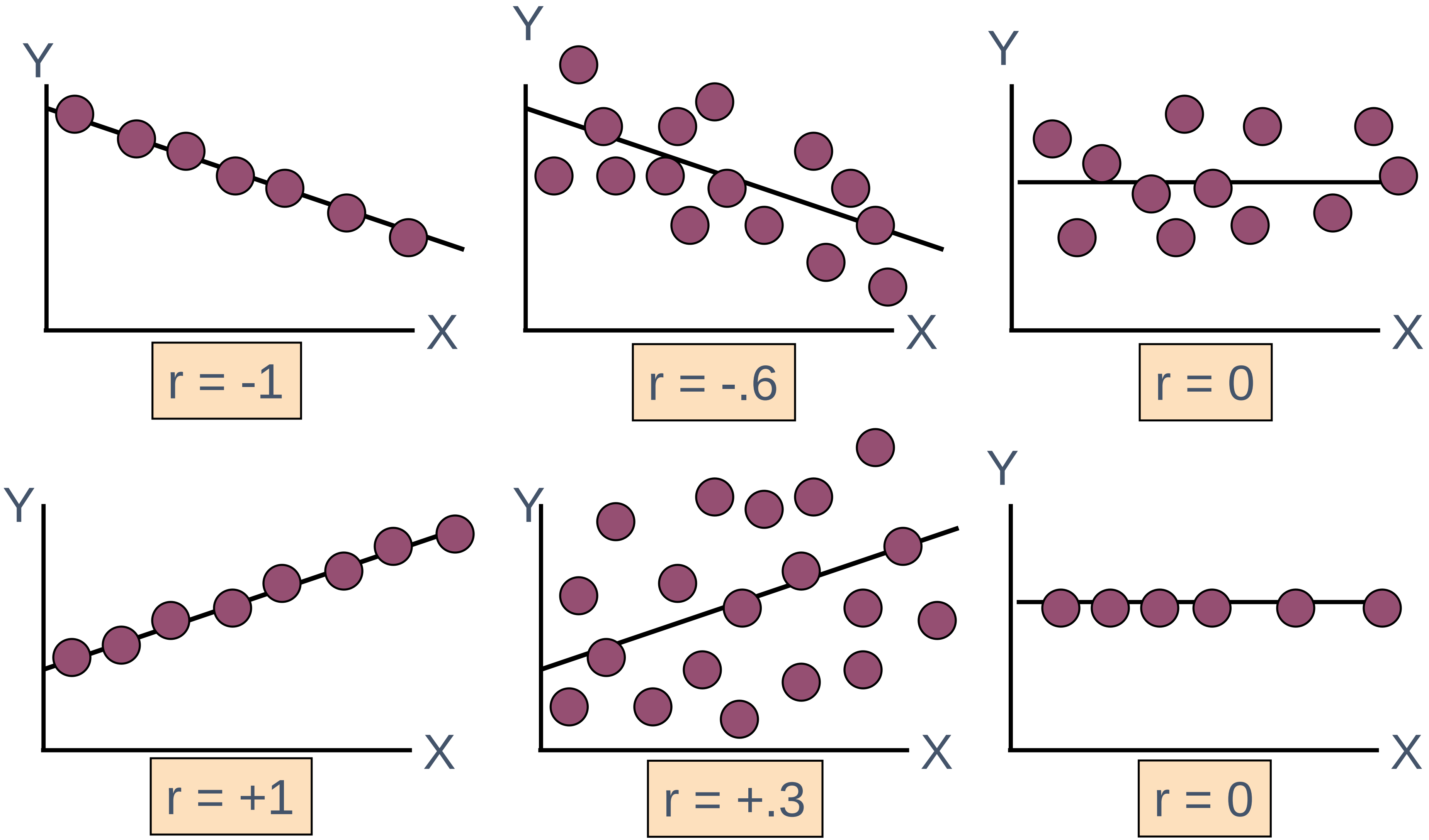
Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

# Correlation

- Measures the relative strength of the *linear* relationship between two variables
- **Unit-less**
- Ranges between **–1 and 1**
- The closer to –1, the stronger the **negative linear** relationship
- The closer to 1, the stronger the **positive linear** relationship
- The closer to 0, the **weaker** any positive linear relationship

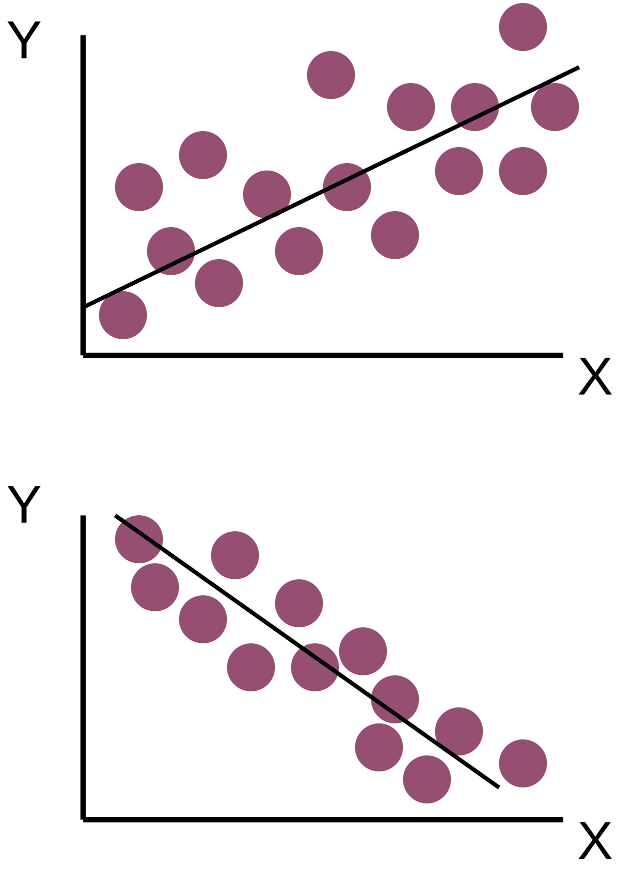
# Scatter Plots of Data with Various Correlation Coefficients



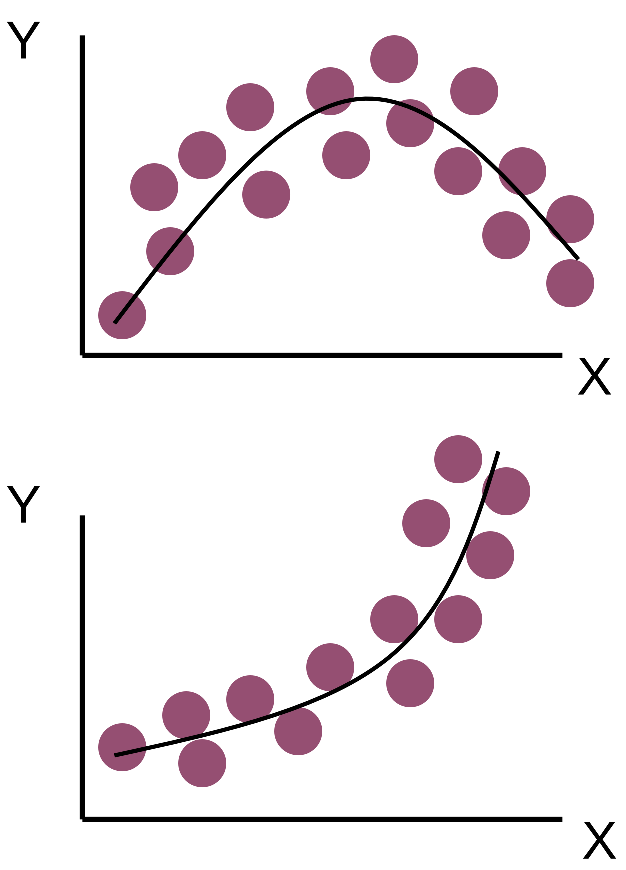

▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

# Linear Correlation

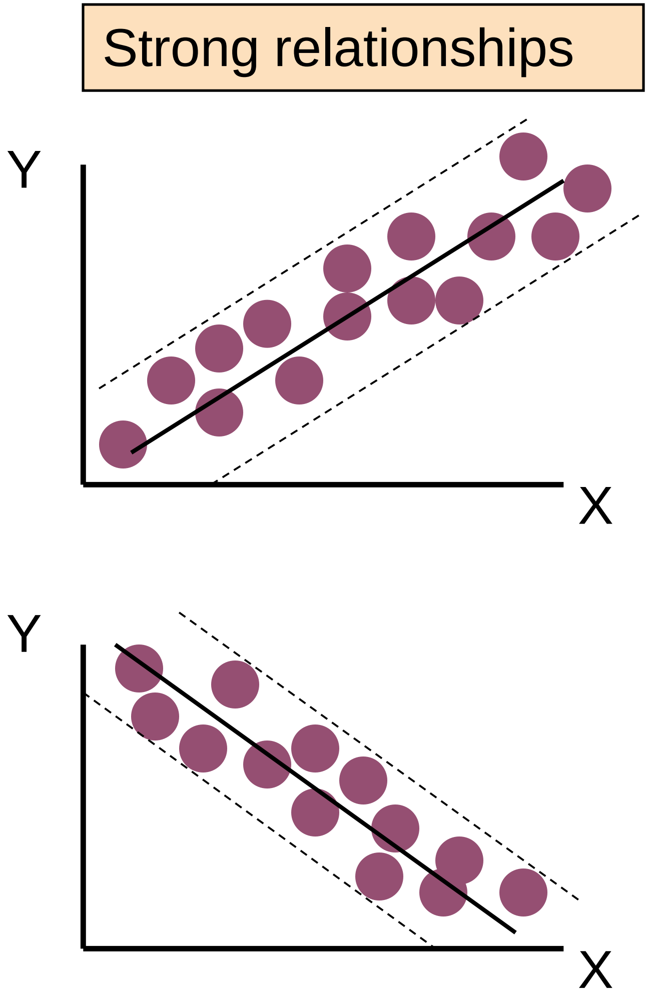
Linear relationships

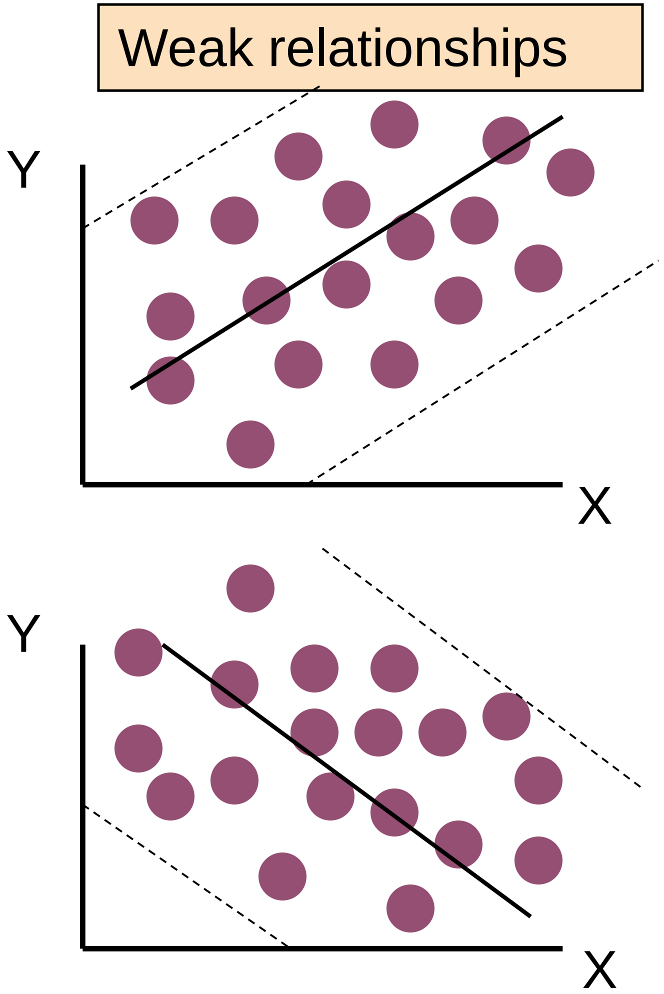



Curvilinear relationships



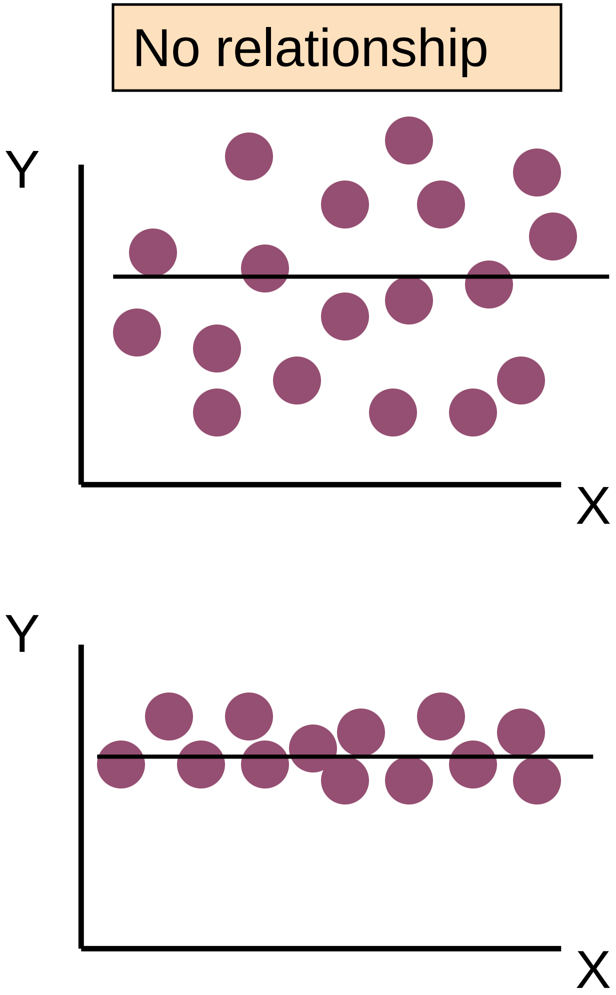
# Linear Correlation



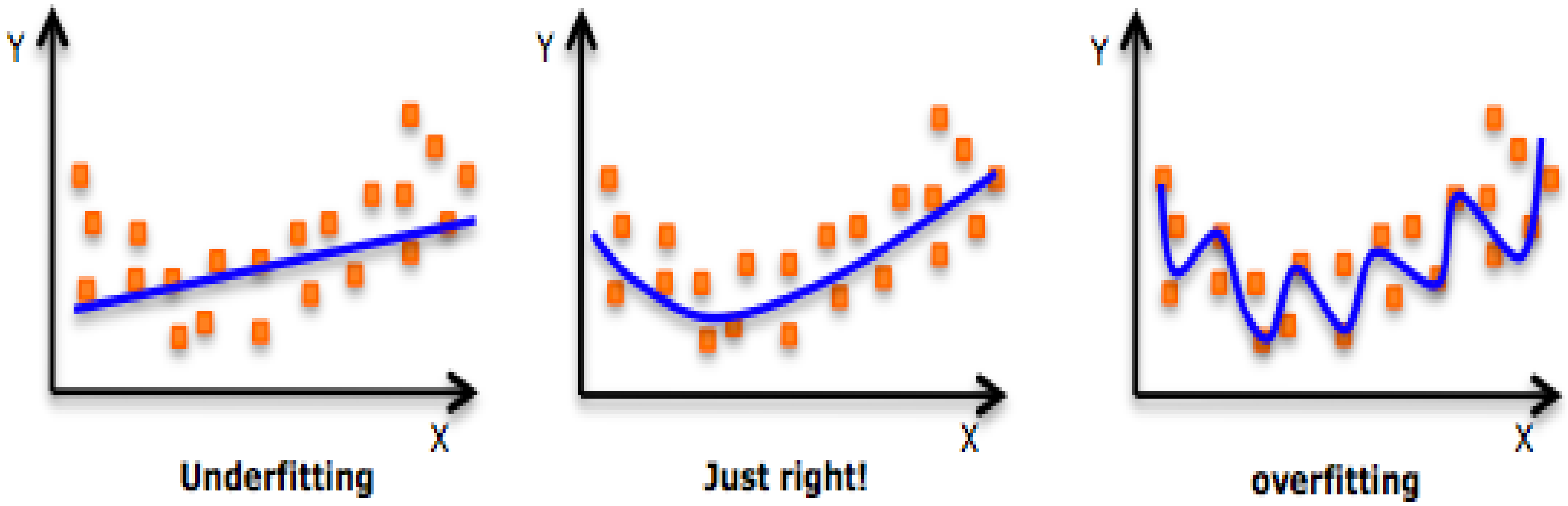




- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

# Linear Correlation




▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall


Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting

Simple Linear Regression

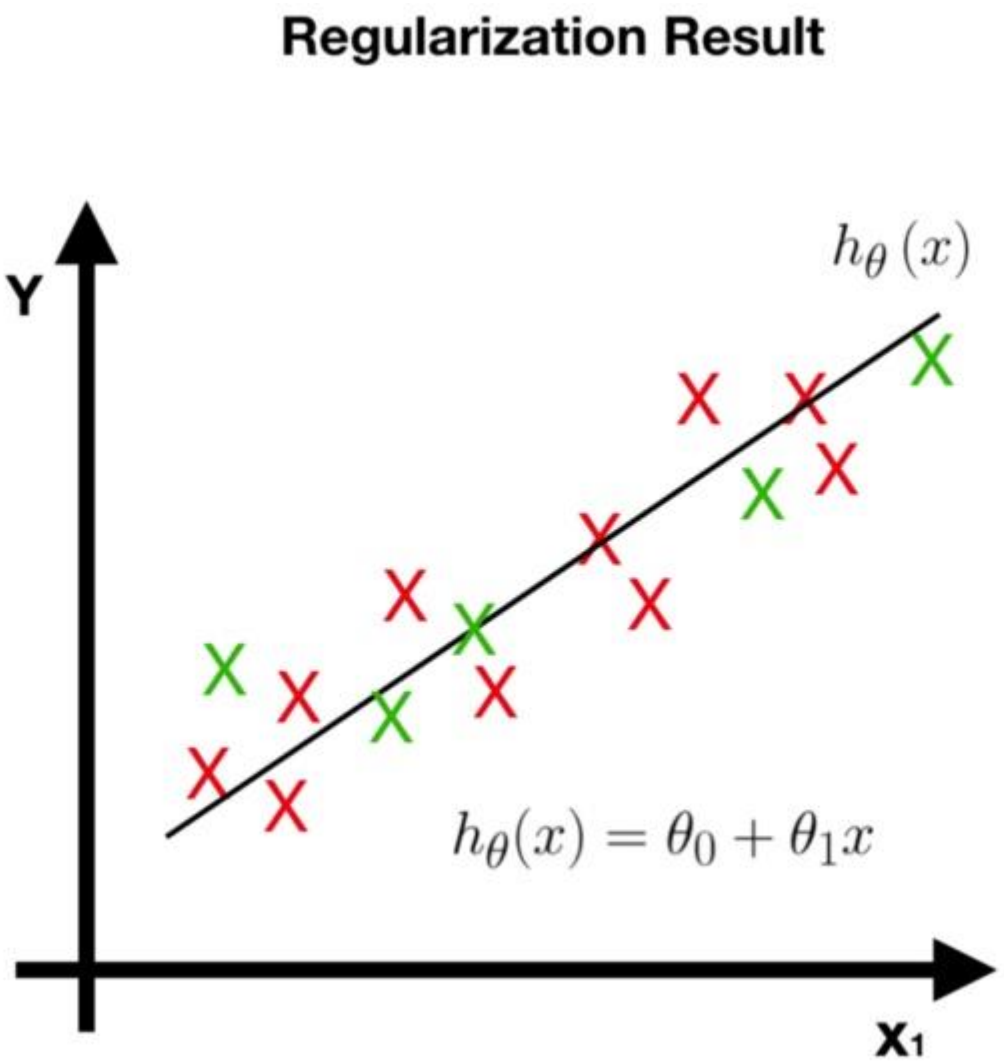
$$y = b_0 + b_1 x_1$$

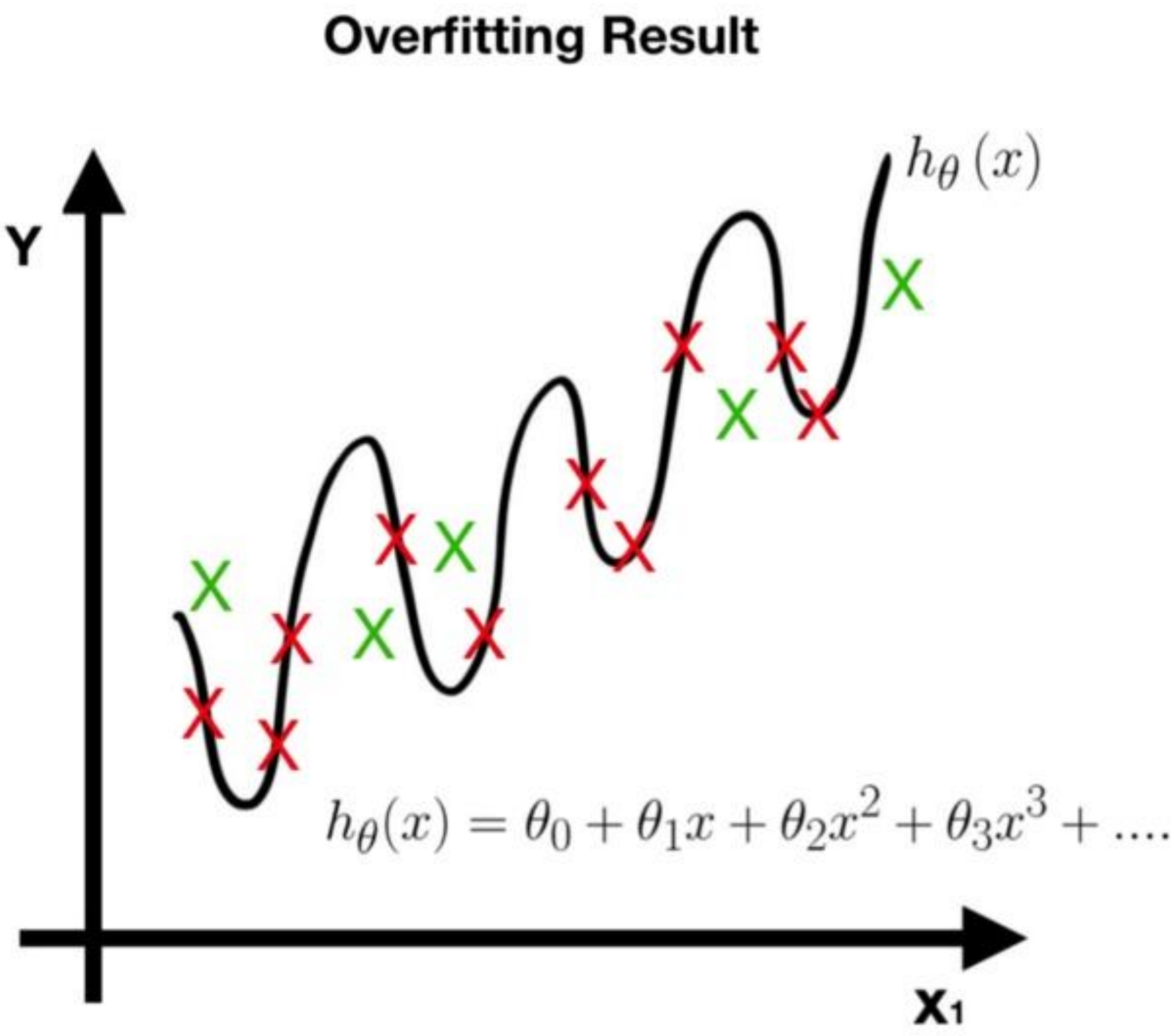
Multiple Linear Regression

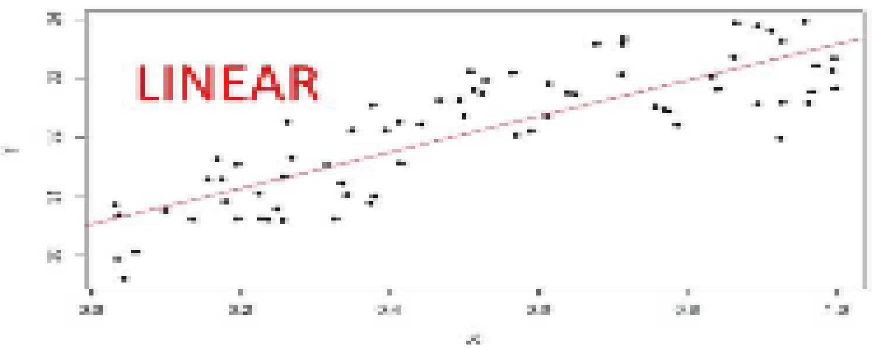
$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$$

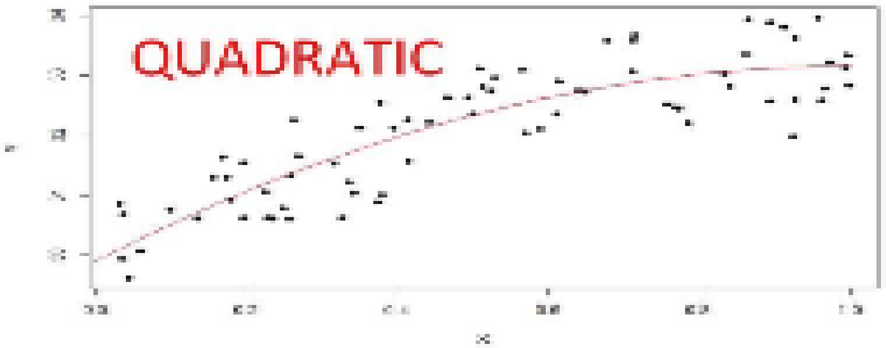

Regularization Result
Y
$h_\theta(x)$
$h_\theta(x) = \theta_0 + \theta_1 x$
$X_1$


Overfitting Result
Y
$h_\theta(x)$
$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + ....$
$X_1$



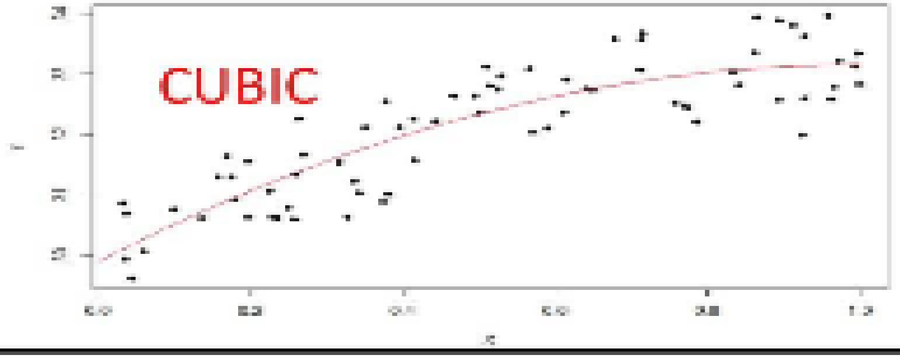

Multiple R-squared: 0.7044

$Y = 30.53 + 3.05*X$



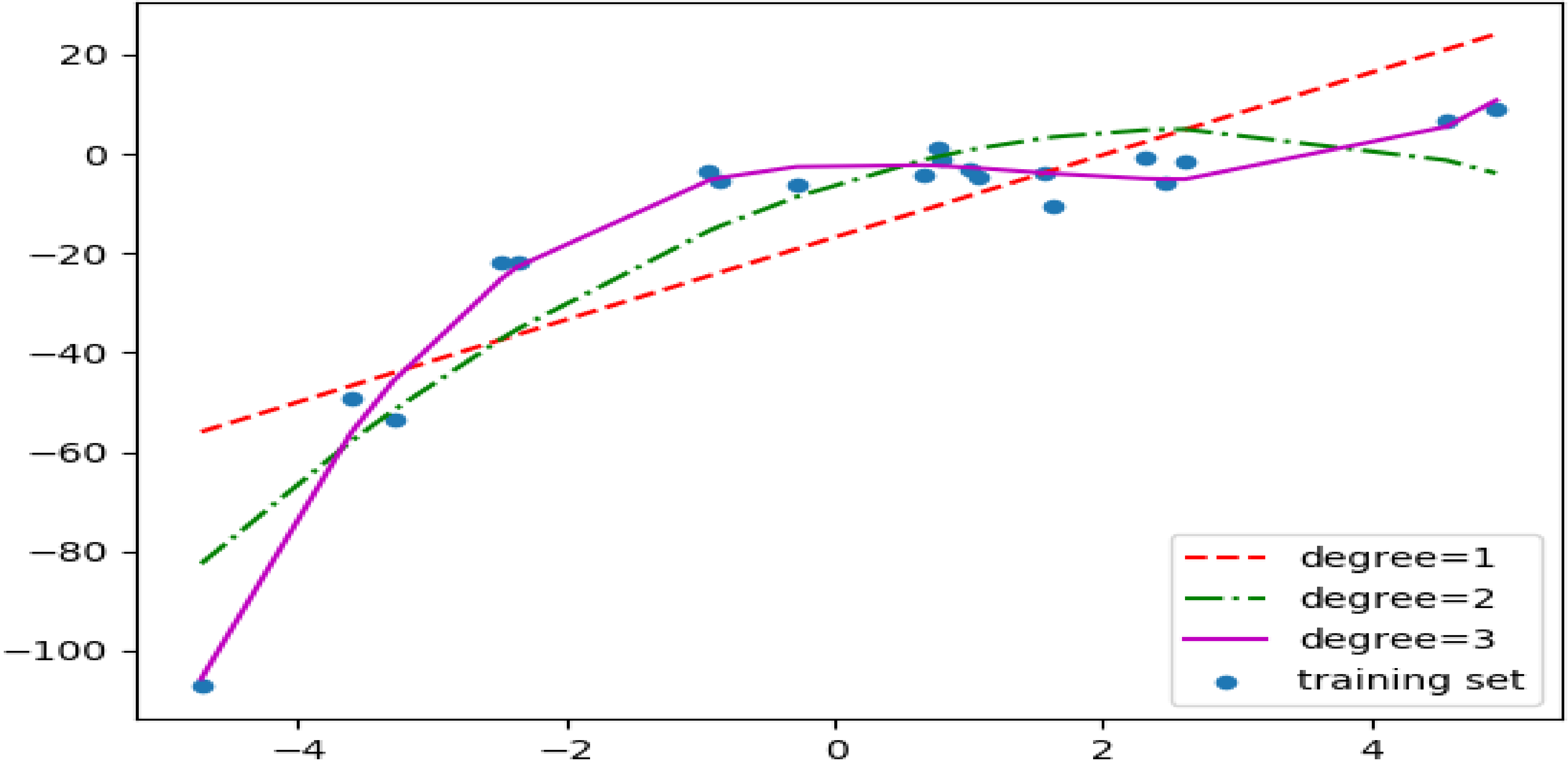

Multiple R-squared: 0.7559

$Y = 29.90 + 6.48*X - 3.22*X^2$




Multiple R-squared: 0.7623

$Y = 30.17 + 3.61*X + 3.71*X^2 - 4.48*X^3$



20
0
−20
−40
−60
−80
−100
−4
−2
0
2
4
degree=1
degree=2
degree=3
training set

# Iris dataset

- Many exploratory data techniques are nicely illustrated with the iris dataset.
  - Dataset created by famous statistician Ronald Fisher
  - 150 samples of three species in genus *Iris* (50 each)
    - *Iris setosa*
    - *Iris versicolor*
    - *Iris virginica*
  - Four attributes
    - sepal width
    - sepal length
    - petal width
    - petal length
  - Species is class label

*Iris virginica*. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.