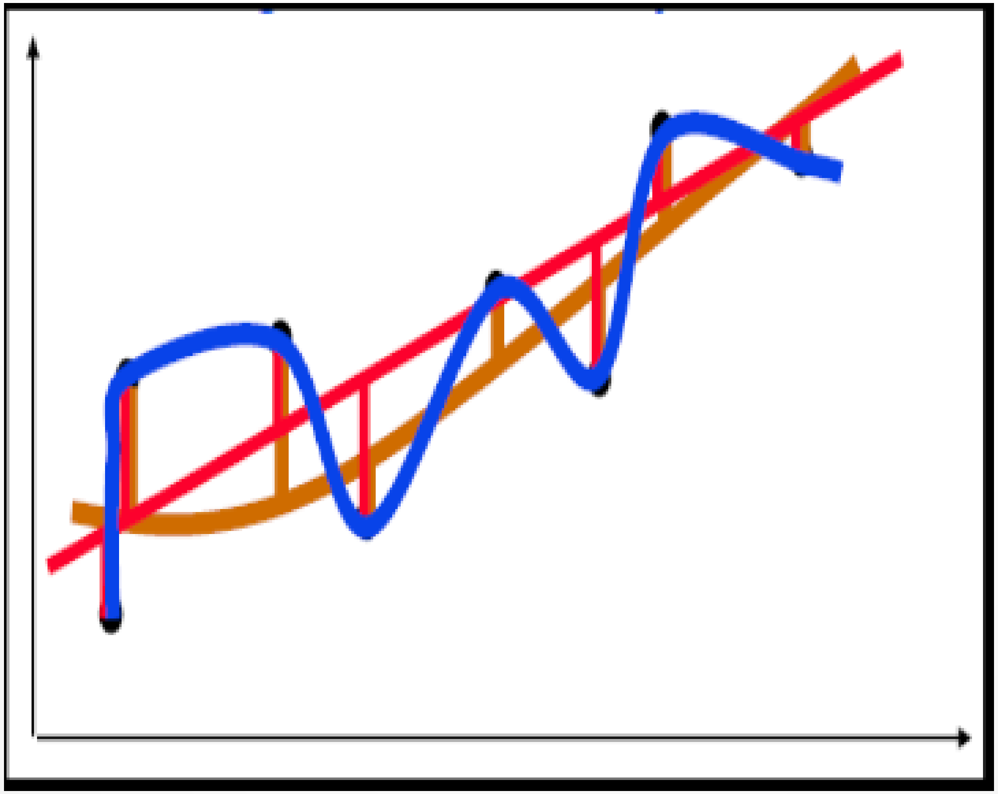
Regression

Jony Sugianto
jony@evolvemachinelearners.com
0812-13086659
github.com/jonysugianto

Regression Analysis

- It is the study of the relationship between variables.
- It is one of the most commonly used tools for business analysis.
- It is easy to use and applies to many situations.



Regression types

- Simple Regression: single explanatory variable
- Multiple Regression: includes any number of explanatory variables.

Regression types

- Linear Regression: straight-line relationship

  Form: $y=mx+b$
- Non-linear: implies curved relationships

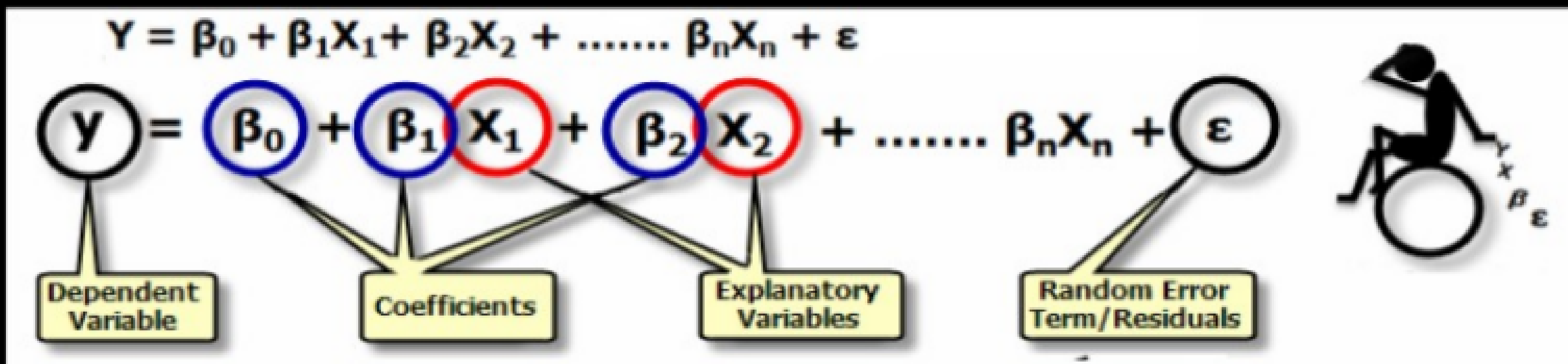
  logarithmic relationships

Regression types

- Cross Sectional: data gathered from the same time period
- Time Series: Involves data observed over equally spaced points in time.

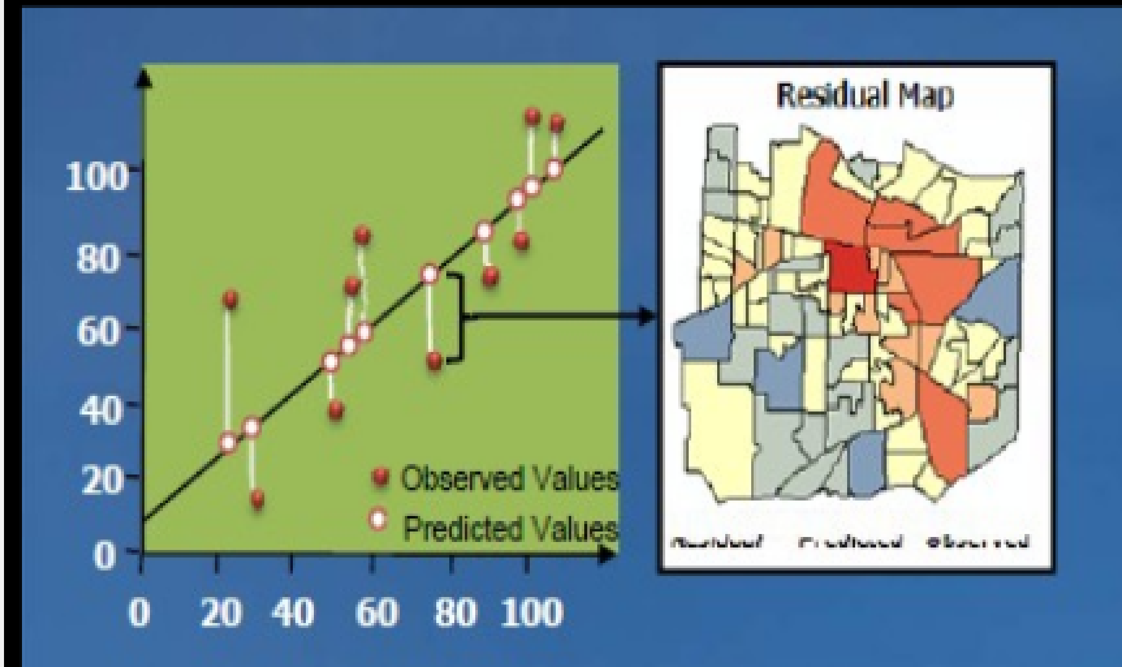
EVOLVE
MACHINE LEARNERS

Vocabulary



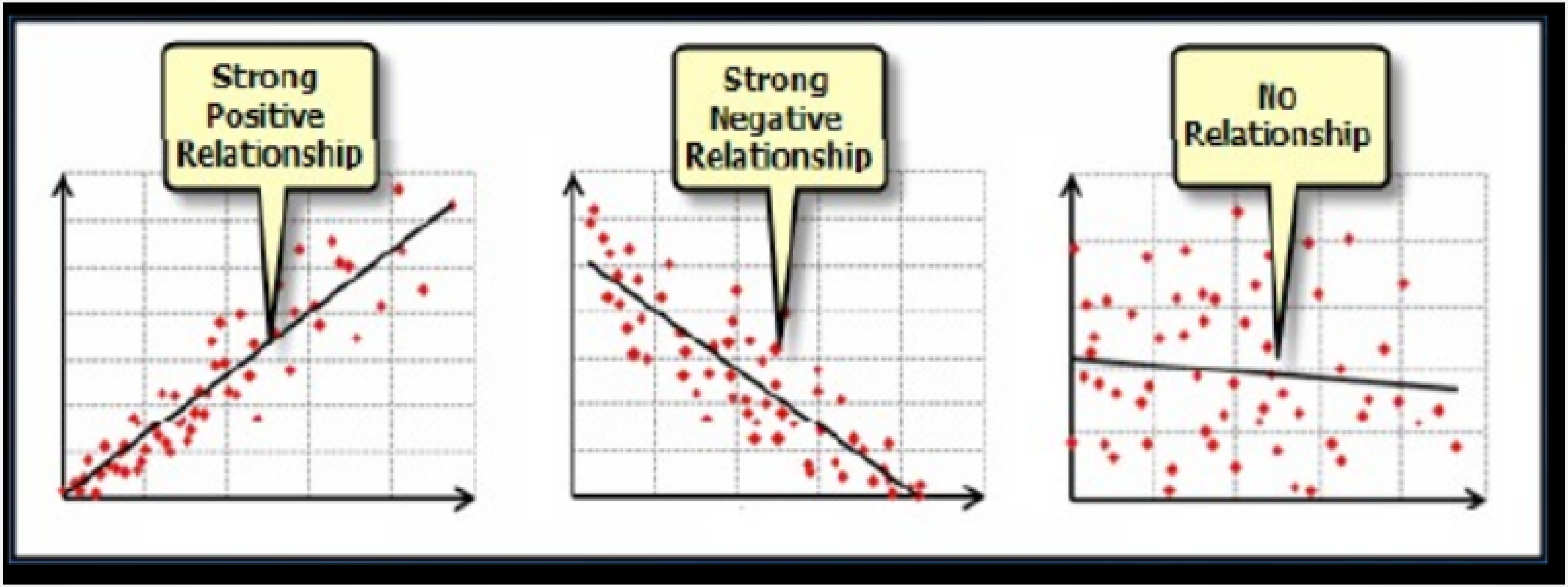
- <u>Dependant variable</u>: the single variable being explained/ predicted by the regression model
- <u>Independent variable</u>: The explanatory variable(s) used to predict the dependant variable.
- <u>Coefficients (β):</u> values, computed by the regression tool, reflecting explanatory to dependent variable relationships.
- <u>Residuals (ε):</u> the portion of the dependent variable that isn't explained by the model; the model under and over predictions.

Simple Linear Regression Model



- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression Model



Ordinary Least Square Method for Linear Regression

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

Solving b0 and b1 using following formulas:

$$\beta_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

In these equations x¯ is the mean value of input variable X and y¯ is the mean value of output variable Y.

Predicting Brain Weight from Head Size

	Gender	Age Range	Head Size(cm^3)	Brain Weight(grams)
0	1	1	4512	1530
1	1	1	3738	1297
2	1	1	4261	1335
3	1	1	3777	1282
4	1	1	4177	1590

Model Evaluation for Regression

RMSE(Root Mean Square Error)

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{1}{m}(\hat{y}_i - y_i)^2}$$

Coefficient of Determination

$$SS_t = \sum_{i=1}^{m}(y_i - \bar{y})^2$$

$$SS_r = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$R^2 \equiv 1 - \frac{SS_r}{SS_t}$$

R^2 Score usually range from 0 to 1. It will also become negative if the model is completely wrong.

Predicting Brain Weight from Head Size

Linear Regression from scratch

Predicting Brain Weight from Head Size

Linear Regression using scikit-learn

Gradient Descent for Multiple Linear Regression

$$Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_n x_n$$

Rewrite The Equation by Introducing x0=1

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_n x_n$$

$$x_0 = 1$$

Rewrite into matrix from

$$Y = \beta^T X$$

Where

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & .. & \beta_n \end{bmatrix}^T$$

$$X = \begin{bmatrix} x_0 & x_1 & x_2 & .. & x_n \end{bmatrix}^T$$

Gradient Descent for Multiple Linear Regression

Define the Hypothesis and cost functions

$$h_\beta(x) = \beta^T x \qquad J(\beta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\beta(x^{(i)}) - y^{(i)})^2$$

Update Rule for gradient descent

$$\beta_j := \beta_j - \alpha\frac{\partial}{\partial\beta_j}J(\beta)$$

After Applying the chain rule for derivative

$$\beta_j := \beta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\beta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

Predicting Writing Score from Math and reading scores

	Math	Reading	Writing
0	48	68	63
1	62	81	72
2	79	80	78
3	76	83	79
4	59	64	62

Predicting Writing Score from Math and reading scores

Multiple Regression from scratch

Predicting Writing Score from Math and reading scores

Multiple Regression with scikit