Modul 3

Supervised Part 1

Data Science Program

Purwadhika
Startup and Coding School

Outline

What is Regression ?

Purpose of Regression

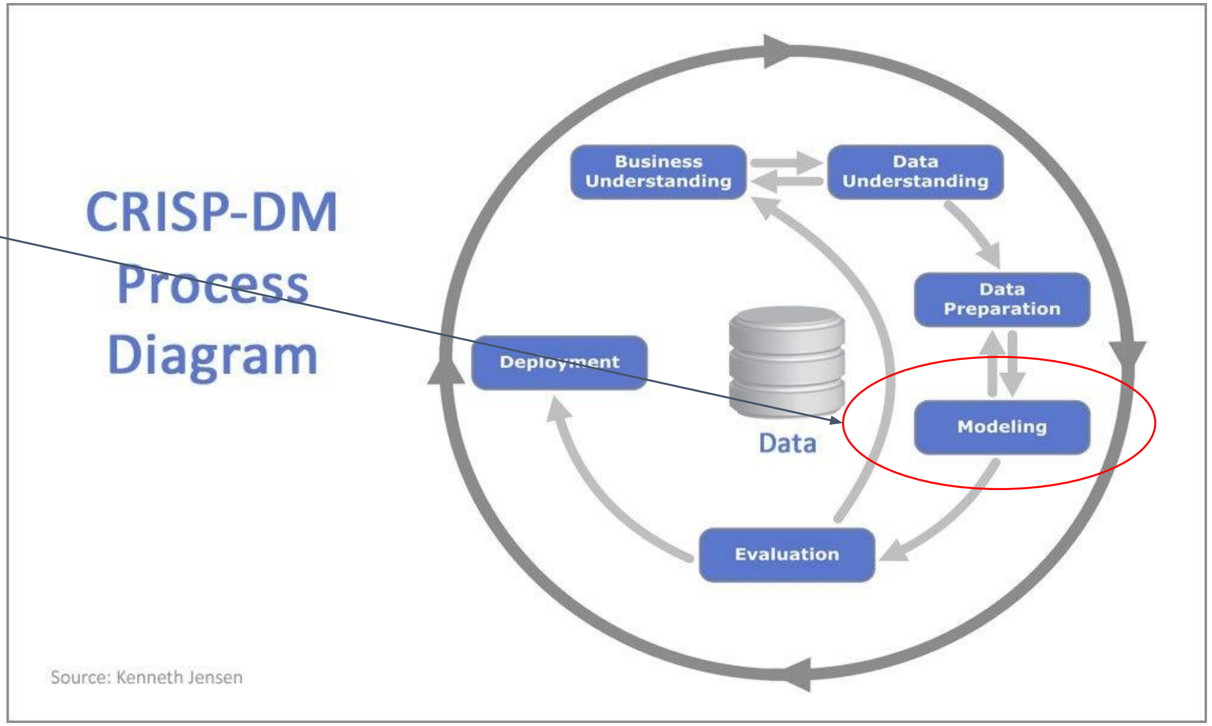
Type of Regression

Simple Linear Regression

Multiple Linear Regression

Diagnostics

Regression with Dummy Variable



Purwadhika
Startup and Coding School

Regression

- Regression is one of method that classified as supervised learning

label = Model function + random error

Y = f(x1, x2, …, xk) + e

How Much???

x1
x2
…
xk

f(x1,...,xk) + error

Y

Linear Regression : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$

What is Regression ?

Num. of Bed	Num. of Room	...	Garage	Pool	House Price
4	10		yes	no	1000M
2	4		yes	no	500M
3	6		no	yes	120M
2	6		no	yes	120M
...	...	...	...	...	...

Houses with known price

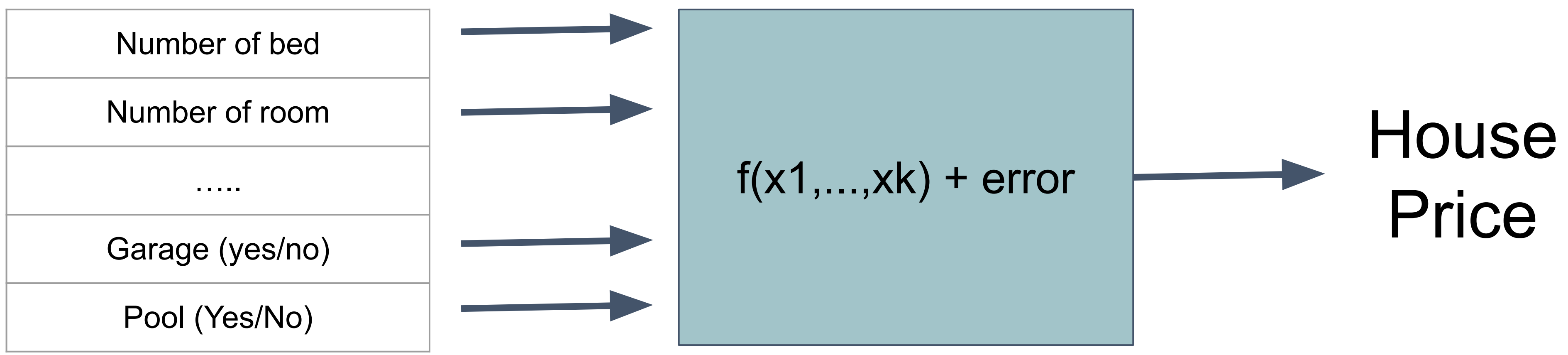
Num. of Bed	Num. of Room	...	Garage	Pool	House Price
4	7		yes	yes	???
2	5		no	no	???
...	...	...	...	...	...

We are interested to predict house with unknown price using the available feature

or

We are interested in analyzing the house price based on its characteristic

House Price



Purpose :
Minimize Overpricing or
Underpricing Phenomenon

Value :
Pricing Strategy

Relationship Review

Do You Still Remember that some event often related to each other

for example:

- air temperature and humidity
- supply and demand
- fertilizer and plant height
- height and weight
- days and COVID-19 victims

There are two types of relationship

- association
- causation

Purwadhika
Startup and Coding School

Response variable and Explanatory Variable Review

When analyzing relationship between two variable usually we must first distinguish between **response variable** (y) and **explanatory variable** (x).

In ML :

- Response Variable → Target or Label
- Explanatory Variable → Feature

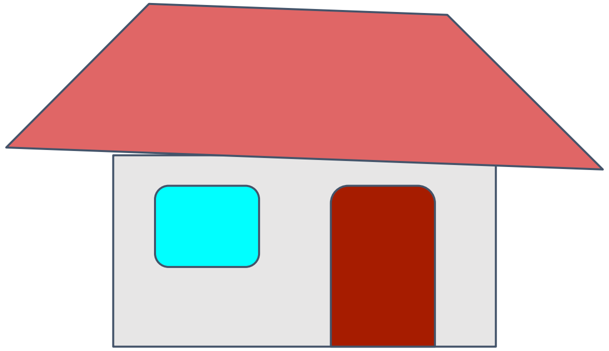
Another Name:

- Response Variable → Dependent Variable
- Explanatory Variable → Independent Variable

Regression Purpose

Prediction

How much for this house ?



Analyze Relationship

how the effect of changes in the number of rooms on the average house price ?

Regression Application Example

Sales Forecasting

Customer Satisfaction

Price Estimation

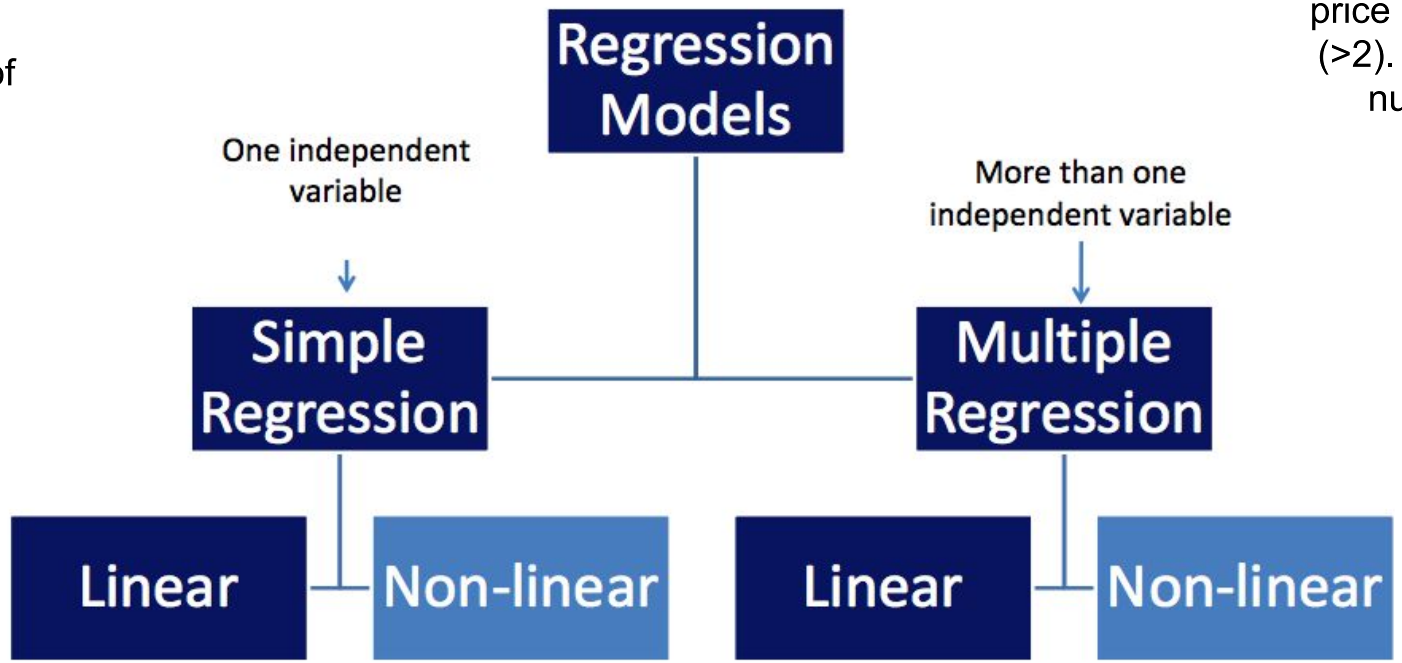
Employment Income

Car CO2 Emission

Types of Regression

SIMPLE:
Predict or analyze house price using only one features. ex number of room

MULTIPLE:
Predict or analyze house price using many features (>2). ex number of room, number of bed, etc



Simple Linear Regression

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \epsilon$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Regression Method

Linear, Polynomial, Lasso, Stepwise, Ridge Regression (Linear)

Poisson Regression (Non-linear)

Decision Tree Regression (Non-parametric)

KNN Regression (Non-parametric)

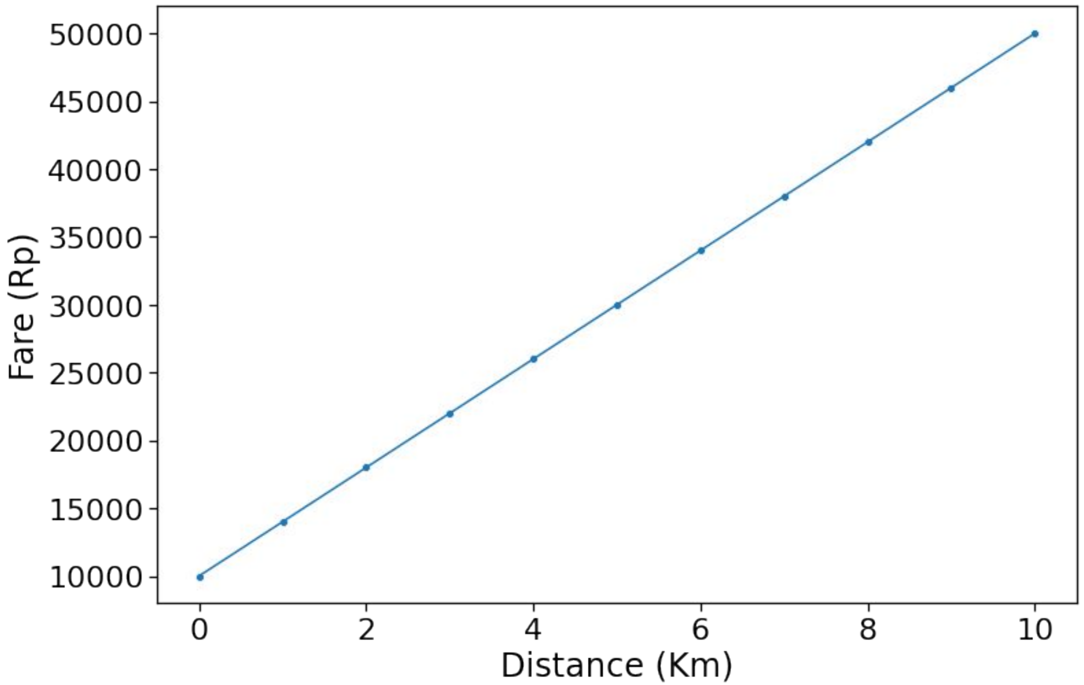
Multivariate Regression (Multilabel)

etc

Simple Linear Regression

Purwadhika
Startup and Coding School

Linear Equation : Taxi Fare (Y) vs Distance (X)



General Linear Equation:

$$Y = a + bx$$

Taxi Fare Linear Equation:

$$Y = 10000 + 4000x$$

Interpretation :

- Slope b = 4000 : For each 1 km the fare will increase Rp. 4,000
- Intercept a = 10000 : This is interpreted as door open rates, when the customer get out of the taxi and the taxi has not been moving at all (x = 0 Km) the customer must pay Rp. 10,000

Purwadhika
Startup and Coding School

Simple Linear Regression Model

- Only one independent variable
- Linear in parameters: linear equation is formed between dependent variable and regression parameters

Population Y-Intercept

Population Slope

Random Error

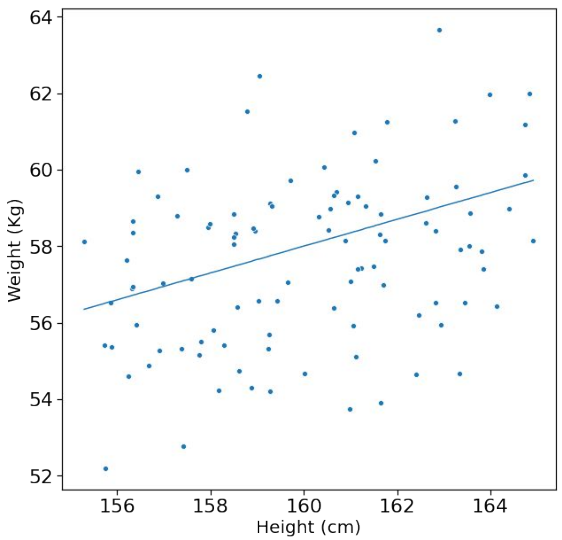
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response)

Independent (Explanatory)

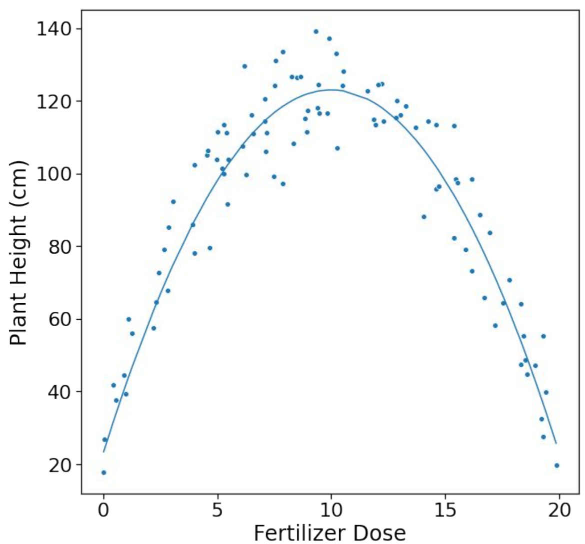
Linear And Nonlinear Relationship

Ex. height and weight



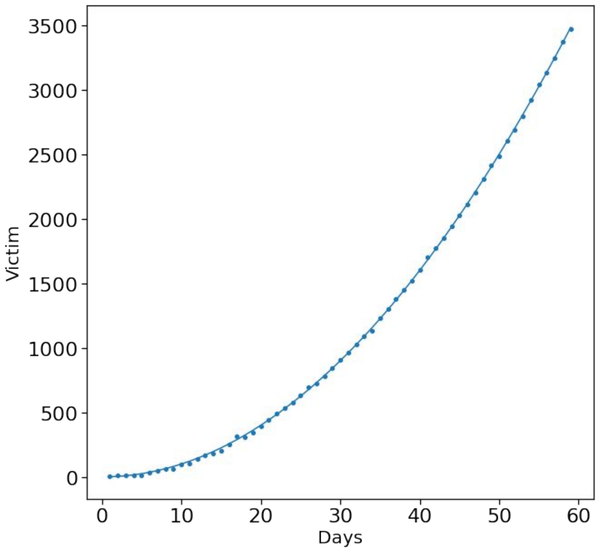
Linear : y = 2 + 0.35x

Ex. fertilizer dose and plant height



Non Linear and Non Monotone

Ex. daily case of COVID-19



Non Linear and Monotone

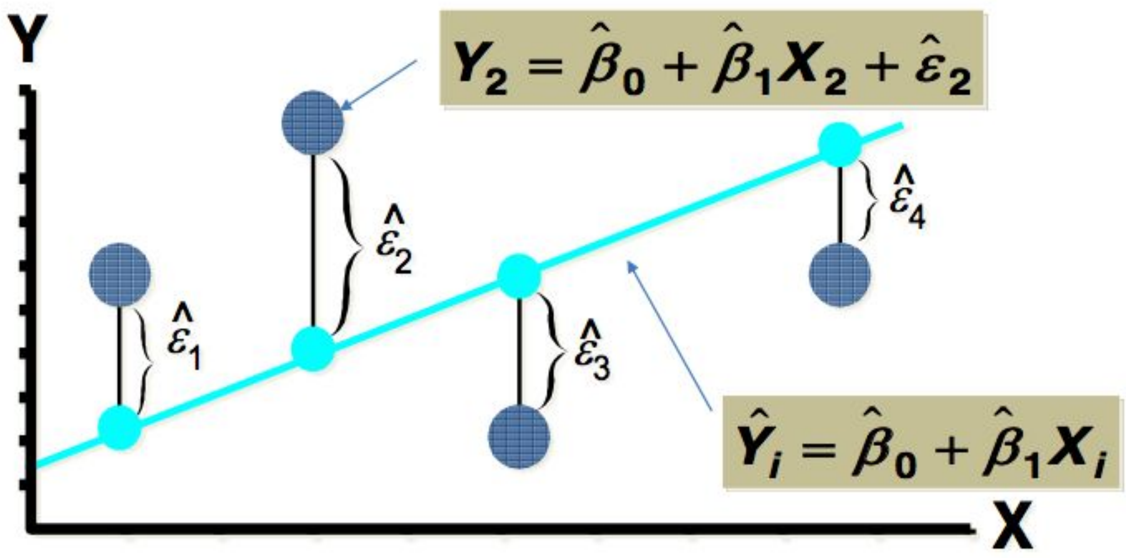
Non Linear Equation Examples

Multiplikatif	$Y = \beta_0 \, x^{\beta_1} \, \varepsilon$
Exponential	$Y = \beta_0 e^{\beta_1 x} \varepsilon$
Reciprocal	$\dfrac{1}{\beta_0 + \beta_1 x + \varepsilon}$

Purwadhika
Startup and Coding School

How To Estimate The Regression Parameters ?

$$\text{LS minimizes} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \ldots + \hat{\varepsilon}_n^2$$



Least Squares (Sum Square Error (SSE)) is a method to estimate regression parameter. Regression parameter estimated by minimizing sum square error.

The are so many method can be used to estimate parameters in linear regression such as:

- resistance line
- weighted least square
- gradient descent
- etc

Purwadhika
Startup and Coding School

Use Case : House Price

Harga Rumah (Rp.juta) (y)	Luas Lantai (m2) (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

- Y → House Price (IDR in millions)
  x → floor area (m2)

- We want to know how floor are can affect house price ?
- we want to know whether the the effect of floor area to house price is significance ?
- How accurate if we use floor area only to predict house price using simple linear regression ?

Purwadhika
Startup and Coding School

Inference in Simple Linear Regression

Interpretation of Simple Linear

F-Test (Simultant Test)

T-Test (Partial test)

Model Performance

Purwadhika
Startup and Coding School

Interpretation Of Regression Parameter

Interpretation depends on the form of mathematical functions used in the model.

Regression Parameter of Linear Model:

- Constant/Intercept : B0
- Coefficient/Slope : B1, B2, …, BK

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Interpretation for linear regression:

The intercept B0, is the mean value of the dependent variable Y, when the independent variable X = 0

The slope Bi, is the change in the value of dependent variable Y, for unit change in the independent variable Xi

* we must carefully interpret the slope because in many case the interpretation only applied within the range of Xi

Purwadhika
Startup and Coding School

House Price (Y) vs Floor Area (X)

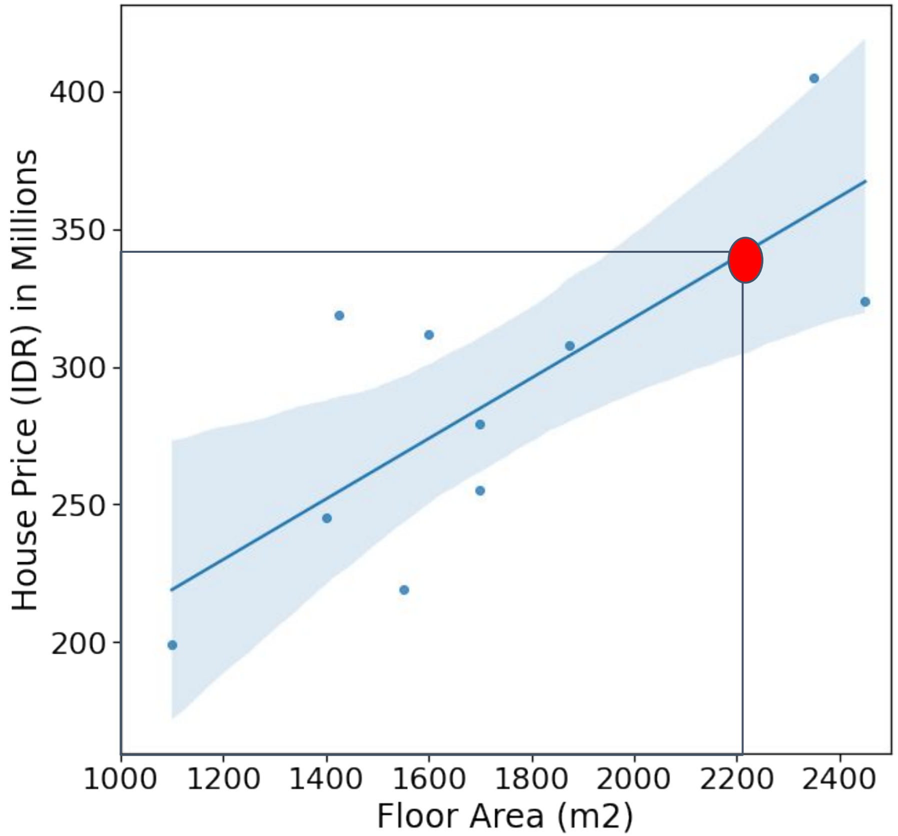
Equation:

$$Y = 98.24 + 0.1098x$$

Interpretation :

- Slope B1 = 0.1098 :
  The house which has a floor area of 2000 m2 is estimated to be more expensive at IDR 21,960,000 compared to a house with a floor area of 1800 m2. We obtain IDR 21,960,000 from 0,1098 x (IDR) 1,000,000 x 200 (m2) = IDR 21,960,000.
  (*This interpretation is only recommended when the height fall between 1100 m2 and 2450 m2)

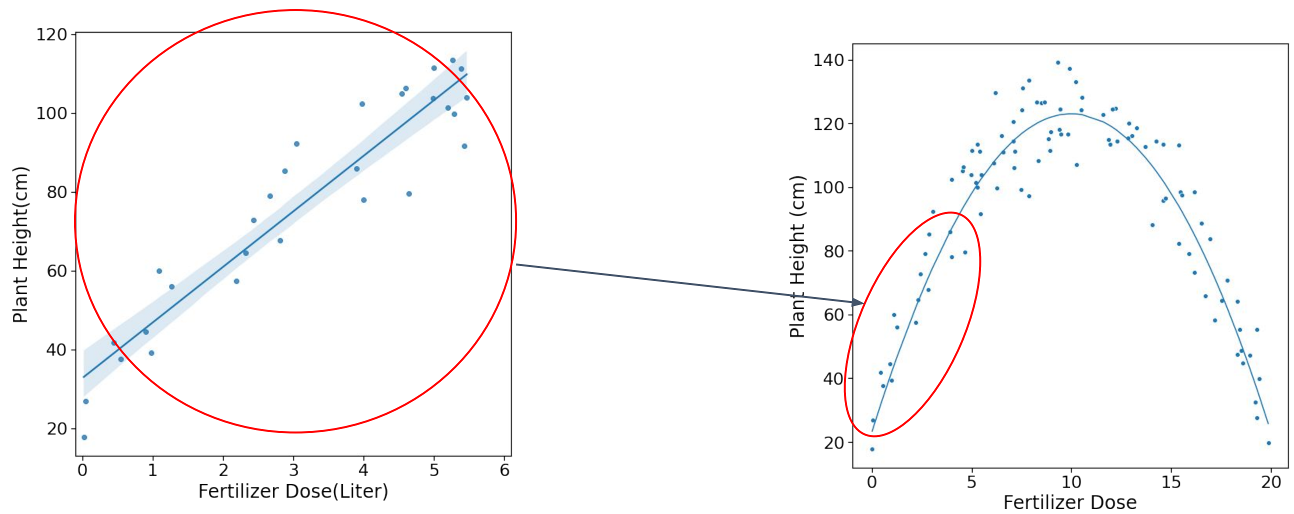
- Intercept B0 = 98.24 : This is not need to be interpreted because there is no house with zero floor area and 0 also fall outside 1100 m2 and 2450 m2 interval.



Purwadhika
Startup and Coding School

Plant Height (cm) (Y) vs Fertilizer Dose (g) (X)

Equation (Linear):
$Y = 32.78 + 14.08x$



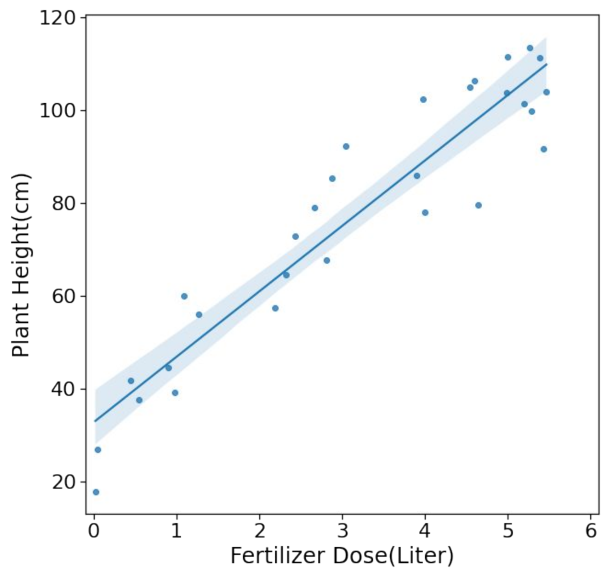
Interpretation (Linear Equation) :

- Slope B1 = 14.08 : When fertilizer dose increase 1 gram the plant height will increase **about** 14.08 cm
  (*This interpretation is only recommended when we give dose between 0 and 10)

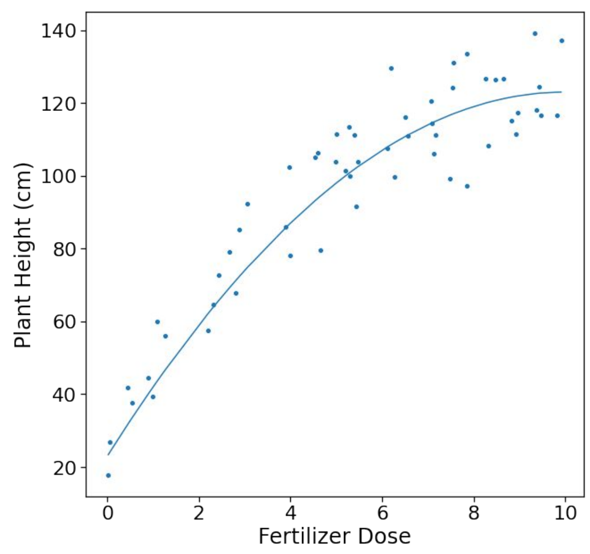
- Intercept B0 = 100 : When we don't give any dose of fertilizer to the plant the plant will grow **about** 32.78 cm

Purwadhika
Startup and Coding School

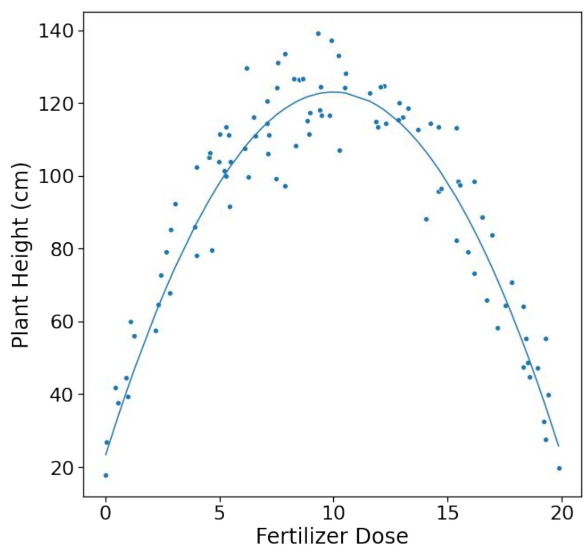
Plant Height (cm) (Y) vs Fertilizer Dose (g) (X)



Fertilizer Dose < 6 gram



Fertilizer Dose < 10 gram



Fertilizer Dose < 20 gram

Purwadhika
Startup and Coding School

ANOVA F-Test for Simple Linear Regression

- In simple linear regression, the F test is used to test whether the independent variable affects the dependent variable.
- The F test requires the assumption that the error normally distributed.
- If the error does not spread normally the test results will not be valid

Hypothesis:

Ho : **B1 = 0**

Ha : **B1 ≠ 0** (two sided only)

Test Statistics : F-Statistics

Rejection Criteria:

P-value ≤ α (two-sided)

Purwadhika
Startup and Coding School

T-Test for Simple Linear Regression : Bo

Hypothesis:

Ho : **B0 = 0**

Ha : **B0** ≠ **0** or **B0** > 0 or **B0** < 0

Test Statistics : t-Student

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

Rejection Criteria:

P-value ≤ α (two-sided)

P-value/2 ≤ α (one-sided)

- The T test in simple linear regression is used to test whether Bo and B1 are significant.
- B0 is tested to infer whether intercept/constant is needed in the model or not.

Purwadhika
Startup and Coding School

T-Test for Simple Linear Regression : B1

Hypothesis:

Ho : **B1** = **0**

Ha : **B1** ≠ **0** or**B1** > 0 or **B1** < 0

Test Statistics : t-Student

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

Rejection Criteria:

P-value ≤ α (two-sided)

P-value/2 ≤ α (one-sided)

- T-Test for B1 have similar function like F-Test.
- T-test for B1 is used to test whether the independent variable affects the dependent variable
- Similar like F-Test but here we can infer the direction as well.

Purwadhika
Startup and Coding School

Regression Model Performance

We want accurate prediction

We can measure a model performance using :

- mse
- rmse
- r2

Residuals = Real - Prediction

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

Purwadhika
Startup and Coding School

MSE and RMSE

- mse and rmse measure how accurate the prediction result
- we want mse and rmse as small as possible
- MSE is the variance of residuals while RMSE is the standard deviation
- mse measure the spread of the residuals.

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

MSE and RMSE Example

Floor Area (m2)	House Price (IDR) in Millions	Predicted House Price (IDR) in Millions	Residuals
1400	245	252.0	-7.0
1600	312	274.0	38.0
1700	279	285.0	-6.0
1875	308	304.0	4.0
1100	199	219.0	-20.0
1550	219	268.0	-49.0
2350	405	356.0	49.0
2450	324	367.0	-43.0
1425	319	255.0	64.0
1700	255	285.0	-30.0

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$$

$$MSE = \frac{(-7)^2 + 38^2 + \ldots + (-30)^2}{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

$$RMSE = \sqrt{\frac{(-7)^2 + 38^2 + \ldots + (-30)^2}{10}}$$

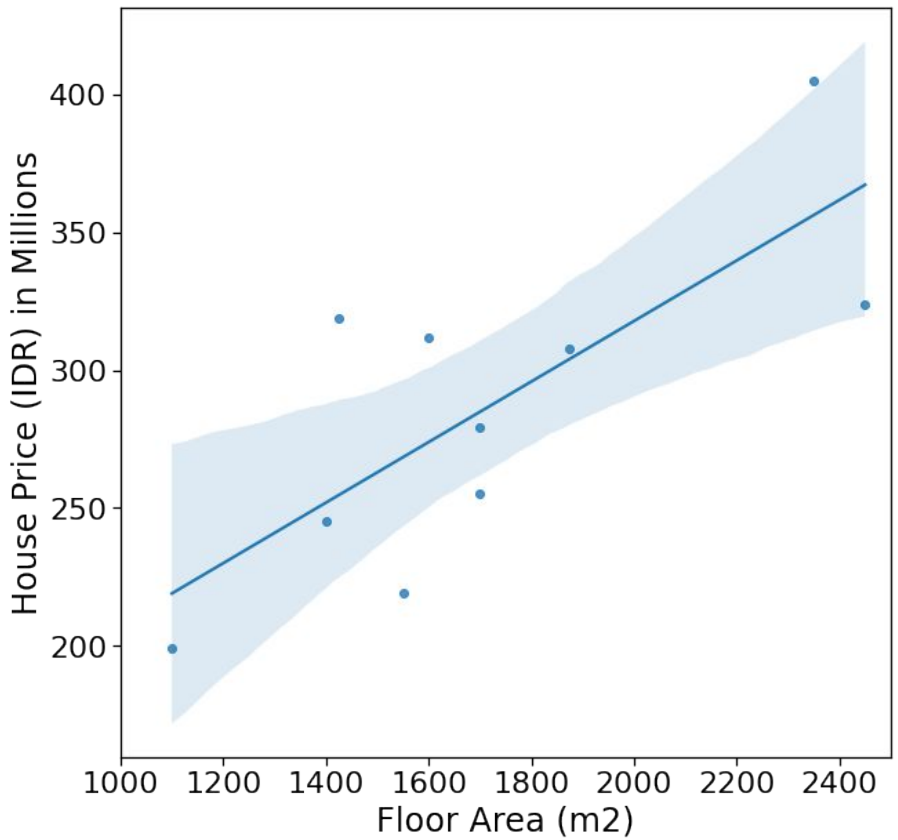
MSE = 1359.2

RMSE = 36.867

R-Square

The goodness of fit of regression equation can be measured using Coefficient Determination (R-Square)

- Coefficient Determination measure how well the regression line fits the data
- Coefficient Determination lies below 1. it's also standardize version of MSE.
- The closer to 1 the better the regression line in representing data.
- interpretation : Percentage of the variation that can be explained by the regression equation



Purwadhika
Startup and Coding School

R Square

Y = Systematic Component (from X) + Non-Systematic Component (error)

SSR

SSE

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

SST = SSR + SSE

- SST : Sum Square Total
- SSR : Sum Square Regression
- SSE : Sum Square Error

Purwadhika
Startup and Coding School

R-Square Analogy

SST

SSR

SSE

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Purwadhika
Startup and Coding School

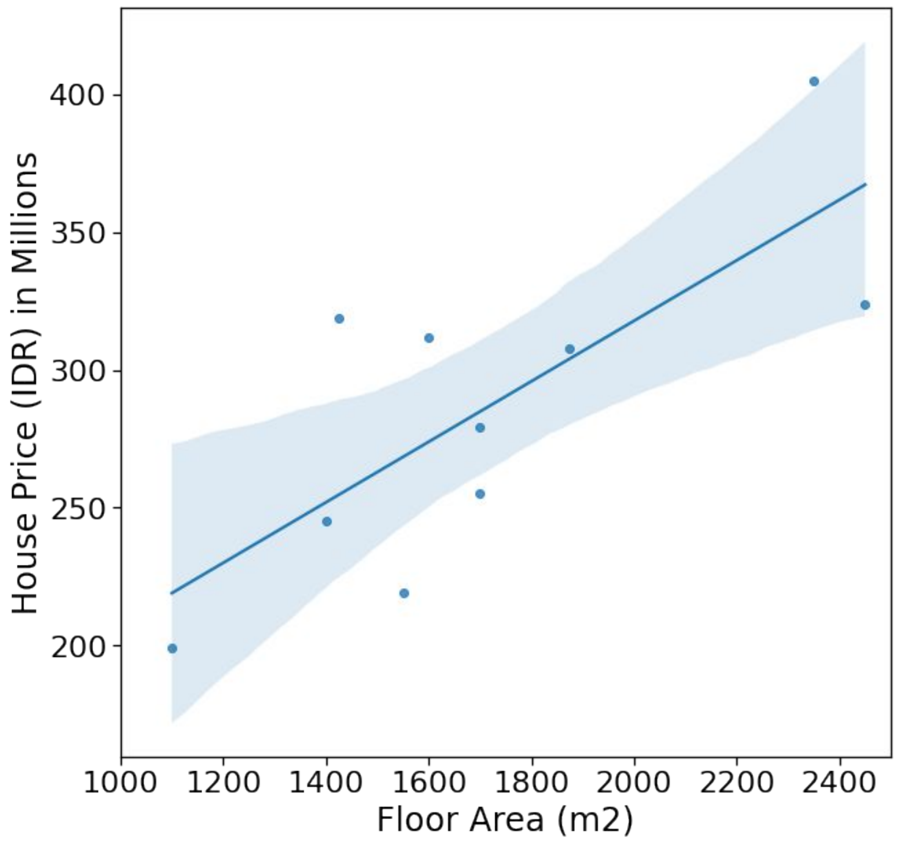
R Square Example

$$SSE = MSE * n = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{1359.2 * 10}{(245 - 286.5)^2 + (312 - 286.5)^2 + \ldots + (255 - 286.5)^2}$$



MSE = 1359.2

R-Square = 58.30%

Simple Linear Regression

Variable x and y has ***Linear*** relationship	Assumption of the world
$y = \beta_0 + \beta_1 x + \varepsilon$, ***Minimize SSE***	Fitting a model
Is x really related to y? ***Is β_1 statistically significant?***	Validating the model
Predict y for a given x.	Using a model

Purwadhika
Startup and Coding School

Use case : Tips Data

Food servers' tips in restaurants may be influenced by many factors, including the nature of the restaurant, size of the party, and table locations in the restaurant. Restaurant managers need to know which factors matter when they assign tables to food servers. For the sake of staff morale, they usually want to avoid either the substance or the appearance of unfair treatment of the servers, for whom tips (at least in restaurants in the United States) are a major component of pay. In one restaurant, a food server recorded the following data on all customers they served during an interval of two and a half months in early 1990. The restaurant, located in a suburban shopping mall, was part of a national chain and served a varied menu. In observance of local law, the restaurant offered to seat in a non-smoking section to patrons who requested it. Each record includes a day and time, and taken together, they show the server's work schedule.

Python Exercise : Simple Linear Regression

Analyze tips data from seaborn

- Total Bill as Independent Variable
- Tips as Dependent Variable

Analyze the relationship

Apply Simple Linear Regression

Perform F Test and T Test

Interpret the result

* use α 5%

Multiple Linear Regression

Purwadhika
Startup and Coding School

Multiple Linear Regression

- Several independent variables may influence the change in dependent variable we are trying to study
- Linear in parameters: linear equation is formed between response variable and regression parameters

Population Y-intercept

Population slopes

Random error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Dependent (response) variable

Independent (explanatory) variables

Inference in Multiple Linear Regression

Interpretation of Multiple Linear Regression

F-Test (Simultant Test)

T-Test (Partial test)

Model Performance

Plant Height (Y) vs Fertilizer Dose and Temperature

Model:

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon}$$

Equation:

Y = 90 + 2x1 + 0.3x2

Y = Plant Height

x1 = Fertilizer Dose (range 0-10 Liter)

x2 = Temperatur (C) (range 30 C - 35 C)

Interpretation :

- B0 = 90 : When we don't give any dose of fertilizer to the plant and the temperature is 30 C the plant will grow **about** 90 + (0.3*30) cm = 90.9 cm

- B1 = 2 : When fertilizer dose increase 1 Liter the plant height will increase **about** 2 cm
  *This interpretation is only recommended when we give dose between 0 and 10 Liter and no changes in another variable (Temperature)

- B2 = 0.3 : When temperature increase 1 C the plant height will increase **about** 0.3 cm
  *This interpretation is only recommended when we give dose between 30 C and 35 C and no changes in another variable (Fertilizer Dose)

Purwadhika
Startup and Coding School

ANOVA F-Test (Simultant) for Multiple Linear

Hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ldots\ldots = \beta_k = 0$$

$$H_A : \text{Not all } \beta \text{ values are zero}$$

Test Statistics : F-Statistics

Rejection Criteria:

P-value ≤ α (two-sided)

- F-test check for overall significance of multiple regression model.
- F-test checks if there is a statistically significant relationship between Y (dependent variable) and any of the independent variables

T-Test (Partial)

Hypothesis:

Ho : **Bi** = **0**

Ha : **Bi** ≠ **0** (two sided)

Bi > 0 or **Bi** < 0 (one sided)

Test Statistics : t-Student

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

Rejection Criteria:

P-value ≤ α (two-sided)

P-value/2 ≤ α (one-sided)

- T-test checks if there is a statistically significant relationship between Y (dependent variable) and each of the independent variables

Purwadhika

Startup and Coding School

Goodness Of Fit Model : Adjusted R-Square

$$R_A^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

R_A^2 = Adjusted R - Square

n = number of observations

k = number of explanatory variables

- SST (Total Sum of Squares):
- SSE (Sum of Squares Error):
- SSR (Sum of Squares Regression):

Purwadhika
Startup and Coding School

Python Exercise : Multiple Linear Regression

Analyze tips data from seaborn

- Total Bill and Size as Independent Variable
- Tips as Dependent Variable

Analyze the relationship

Apply Multiple Linear Regression

Perform F Test and T Test

Interpret the result

* use α 5%

Residual Analysis

Purwadhika
Startup and Coding School

What is residual ?

x1	x2	Y	Predictions (Y = 90 + 2x1 + 0.3x2)	Residuals
3.5	31	107	106.3	0.7
4.1	32	106	107.8	-1.8
6.5	33	109	112.9	-3.9
5	35	112	110.5	1.5

Residuals = Real - Prediction

Equation:

Y = 90 + 2x1 + 0.3x2

Y = Plant Height

x1 = Fertilizer Dose (range 0-10 Liter)

x2 = Temperatur (C) (range 30 C - 35 C)

Why do we need to analyze residuals ?

Ordinary Least Square (OLS) is the most common estimation for Linear model, the estimation have some assumption requirements to be fulfilled in order to get the best estimation.

When we talk about error term, it is the residual instead of the population error term. We need to use the residual as the population error term is unknown.

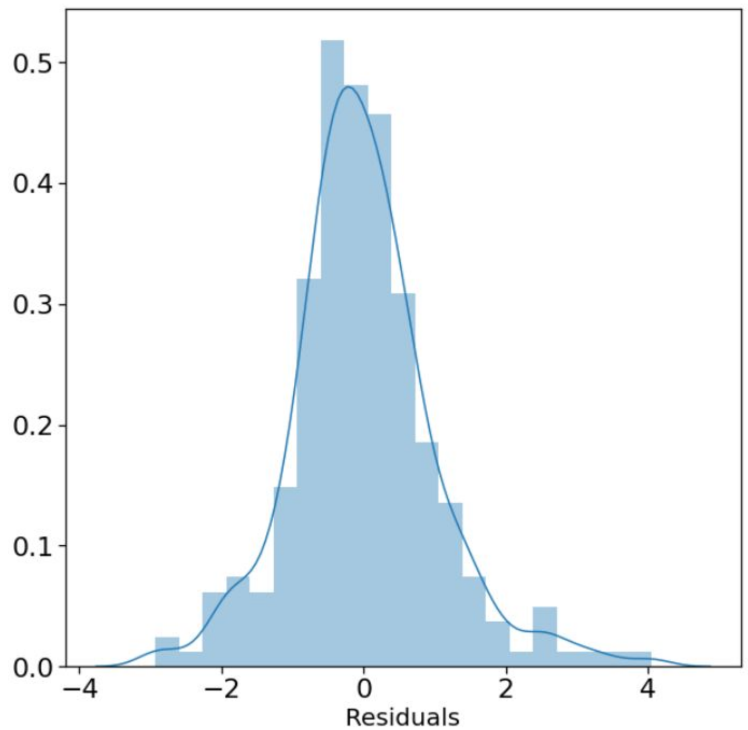
- Residuals → Sample
- Error → Population

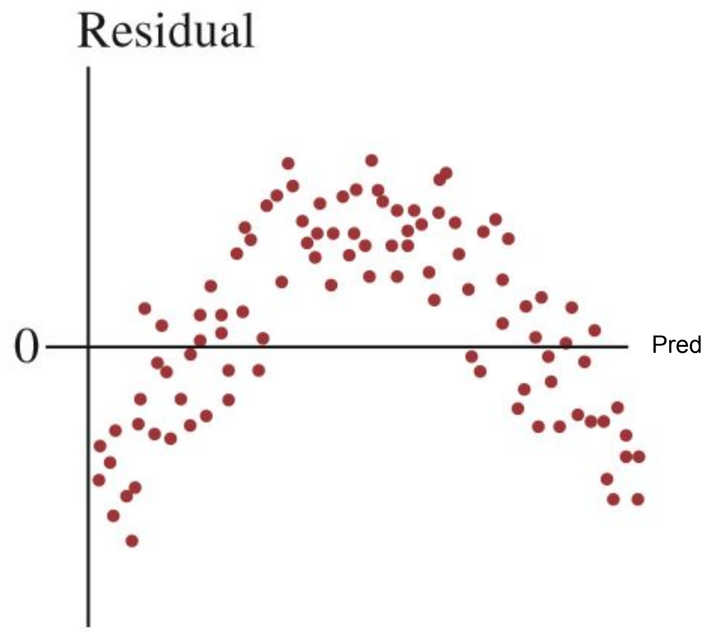
Residuals is used to assess some assumption in model. Each assumption that is violated has its own impact on the results of analysis and predictions.

- The regression model is linear in the parameters and the error term
- Gauss-Markov (Specific to Least Square):
  - The error term has a population mean of zero
  - Observation of the error terms are uncorrelated with each other
  - The error term has constant variance (Homoscedasticity)
- The error term are normally distributed

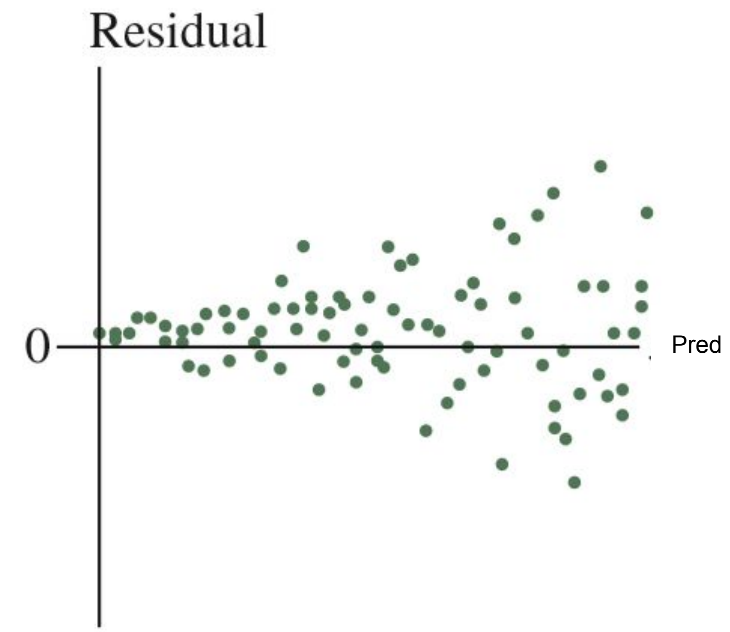
Purwadhika
Startup and Coding School

Residual Analysis





(b) Nonlinear relationship



(c) Nonconstant standard deviation

Purwadhika
Startup and Coding School

Residual Analysis : Normality Assumption

Simultant Test (F-Test) and Partial Test (T-Test) needs normality assumption so the test result more valid.

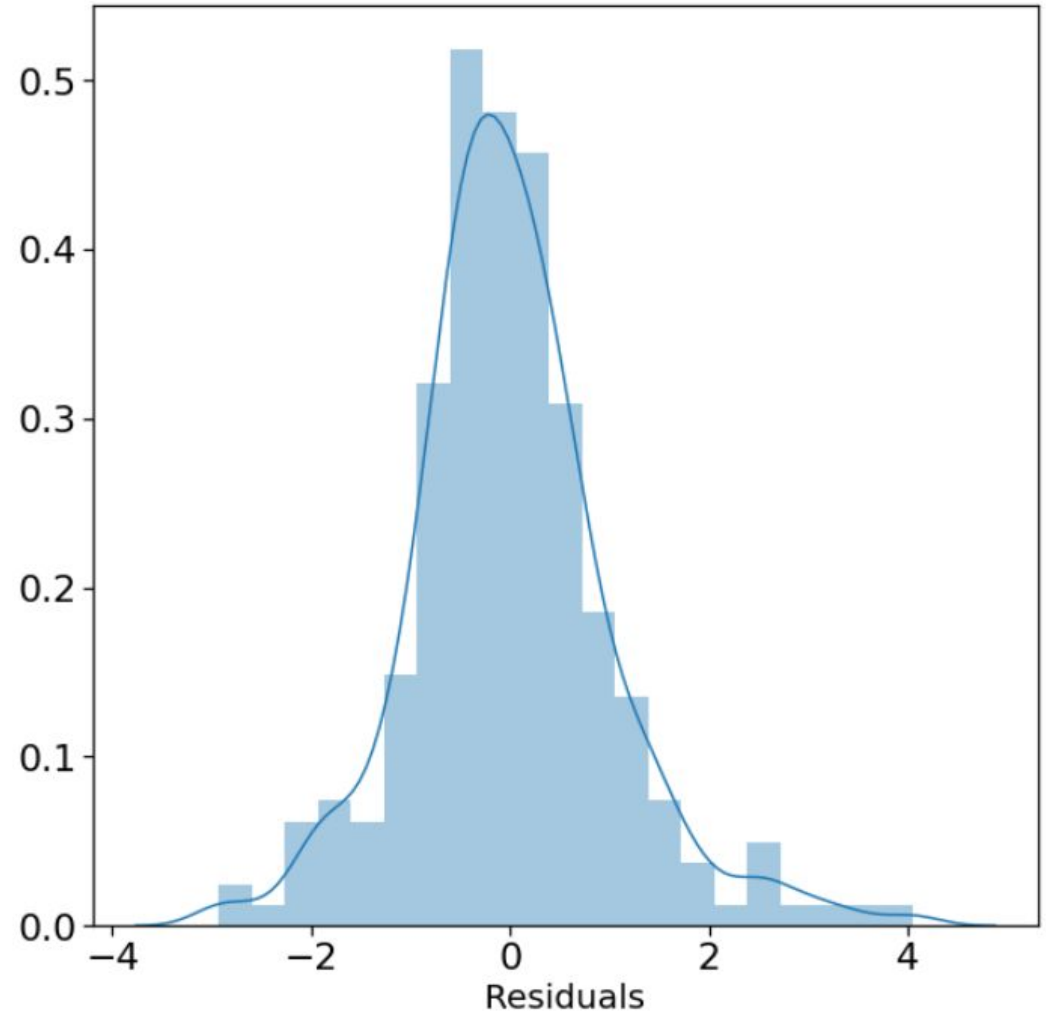
Assessing normality assumption :

1. histogram
2. qqplot
3. normality test such as : kolmogorov-smirnov (KS), Jarque-Bera (JB), etc

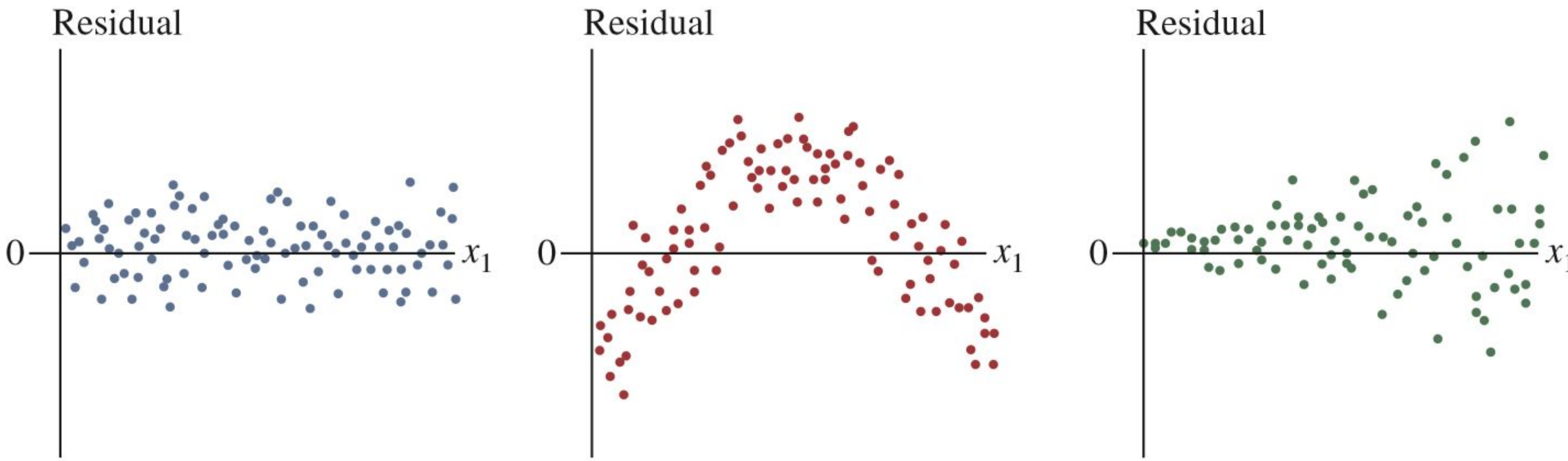
   Ho : Normal

   H1 : Not Normal

   Rejection : P-Value < α



Residual Analysis : Gauss-Markov



(a) Assumptions satisfied

(b) Nonlinear relationship

(c) Nonconstant standard deviation

- The error term has a population mean of zero
- Picture a : all gauss-markov assumption meet
- Observation of the error terms are uncorrelated with each other (picture b : violation of this assumption)
- The error term has constant variance/Non Constant standard deviation (Homoscedasticity) (picture c : violation of this assumption)

Purwadhika
Startup and Coding School

Residual Analysis : Gauss-Markov

<table>
<tr><th>Assumption</th><th>Violation Consequences</th><th>How to Handle Them</th></tr>
<tr><td>Error term mean equal to zero</td><td>Bias in prediction result and regression parameter estimate</td><td rowspan="3">Add nonlinear component
or
Change the model to nonlinear
Change the method
Transformation :
Y-new = Log(Y)
Y-new = sqrt(Y)
etc</td></tr>
<tr><td>Uncorrelated Error</td><td>Simultant Test and Partial Test always tend to reject Ho even when actually there is no relationship,
Overly optimistic R-Square or R-Square adjusted</td></tr>
<tr><td>Constant Variance</td><td>Unstable in prediction result and regression parameter estimate</td></tr>
</table>

Multicollinearity

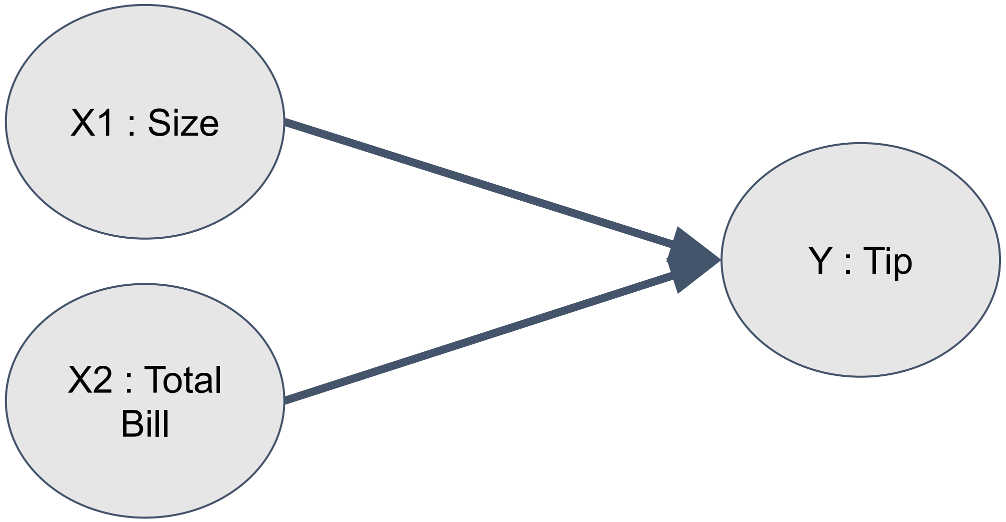
What is Multicollinearity ?

- Existence of high correlation between independent variable is called multi-collinearity
- correlation between independent variable is always existing ;it is just the matter of degree(how strong the correlation)
- We assume that there is no perfect correlation between independent variables when the least square method is used to estimate regression parameters.
- When collinearity exist, interpretation of the slope in multiple linear regression doesn’t apply because the dependent variable is not the only variable that changed.

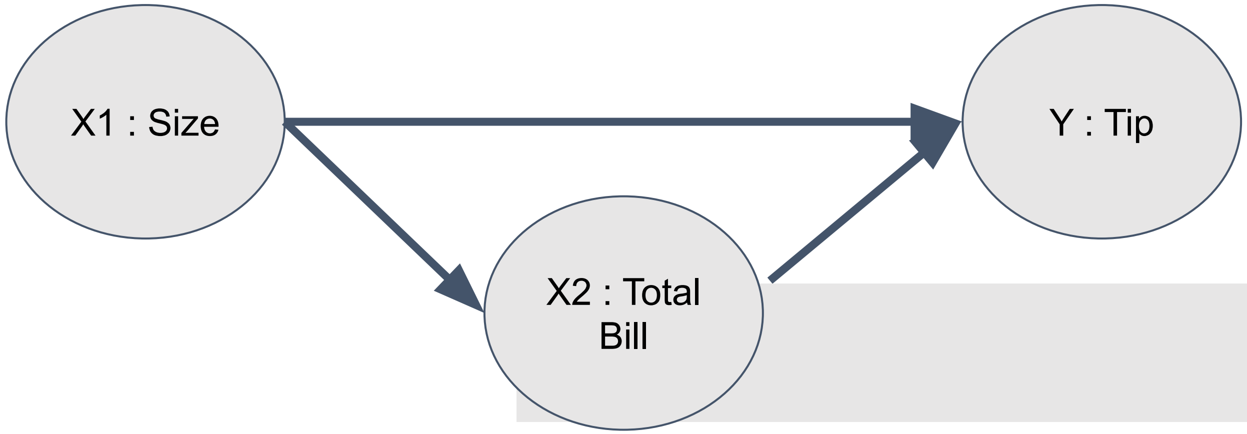
Purwadhika
Startup and Coding School

size	total_bill	tip
2	16.99	1.01
3	10.34	1.66
3	21.01	3.50
2	23.68	3.31
4	24.59	3.61
...	...	...
3	29.03	5.92
2	27.18	2.00
2	22.67	2.00
2	17.82	1.75
2	18.78	3.00

Regression Equation : Y = 0.6689 + 0.1926x1 + 0.0927x2



In Reality :



Why is Multicollinearity Dangerous ?

Danger	**Consequences**
The variances of regression coefficient estimators are inflated	Reliability of Partial Test - biased result of p-value
Adding or removing variables produce large changes in the coefficient estimates	Reliability of Coefficient Regression Interpretation - unstable coefficient
Regression coefficient may have opposite sign	Reliability of Coefficient Regression Interpretation - misinterpretation

Purwadhika
Startup and Coding School

The Characteristics Multicollinearity

- Having High R^2 but only few significant t ratios.

- F-test rejects the null hypothesis, but none of the individual t-tests are rejected.

- Correlations between pairs of X variables (independent variables) are stronger than Correlations between each of X variables with Y variables (dependent variables).

Purwadhika
Startup and Coding School

How to Identify Multicollinearity ?

- The variance inflation factor (VIF) is a relative measure of the increase in the variance in standard error of beta coefficient because of collinearity.
- A VIF greater than 10 indicates that collinearity is very high. A VIF value of more than 4 is not acceptable.

Variance inflation factor associated with introducing a new variable X_j is given by:

$$\mathrm{VIF}(X_j) = \frac{1}{1 - R_j^2}$$

R_j^2 is the coefficient of determination for the regression of X_j as dependent variable

The standard error of the corresponding Beta is inflated by $\sqrt{VIF}$

Purwadhika
Startup and Coding School

Python Exercise : Diagnostics And Multicollinearity

Analyze tips data from seaborn

- Total Bill and Size as Independent Variable
- Tips as Dependent Variable

Apply Multiple Linear Regression

Check The Normality Assumption

Check The Gauss-Markov Assumption

Check The Multicollinearity

Regression With Dummy Variable

What is Dummy Variable ?

In Regression and usually any kind of modeling, categorical variable in Regression model cannot be integrated as it is, because they are not numerical

Gender	Domicile	Age	Income(Y)
Male	Jakarta	34	20M
Female	Bogor	28	15M
Female	Bogor	23	7M
Male	Bekasi	26	9M
Female	Bekasi	29	12M
Female	Jakarta	25	11M
Male	Bekasi	25	9M

Purwadhika
Startup and Coding School

How to Define Dummy Variable ?

Gender
Male
Female
Female
Male
Female

Dummy Gender
1
0
0
1
0

City
Jakarta
Bogor
Bogor
Bekasi
Bekasi

Bogor	Jakarta
0	1
1	0
1	0
0	0
0	0

- When there are k categories, dummy that needed to be defined only k-1 and you may choose which want depend on what you prefered. ex (Jakarta, Bogor, Bekasi) → only Bogor and Bekasi
- if we made k dummy variable, it will lead to collinearity problem in linear regression

Purwadhika
Startup and Coding School

How is The Model Would Be ?

Model :

Income = B0 + B1 Age + B2 Dummy Gender + B3 Bogor + B4 Jakarta + e

Regression Equation :

Income = -19000000 + 1000000 Age + 120000 Dummy Gender + 2300000 Bogor + 2400000 Jakarta

Interpretation :

- B1 = IDR 1,000,000 : When age increase 1 year the income will increase **about** 1M
  *This interpretation only recommended when the age fall between 23 and 34 year and no changes in another variable (Gender, City)
- B2 = IDR 120,000, the average salary for men is higher around IDR 120,000 than the average salary for women
- B3 = IDR 2,300,000, the average salary of people living in Bogor is higher around IDR 2,300,000 than people living in Bekasi
- B4 = IDR 2,400,000, the average salary of people who live in Jakarta is higher around IDR 2,400,000 than people who live in Bogor

Python Exercise : Dummy Variable

Analyze tips data from seaborn

- Total Bill and Size as Numerical Independent Variable
- Sex, smoker, day, and time as Categorical Independent Variable
- Tips as Dependent Variable

Analyze the relationship

Apply Multiple Linear Regression with dummy variable

Perform Simultant Test and Partial Test

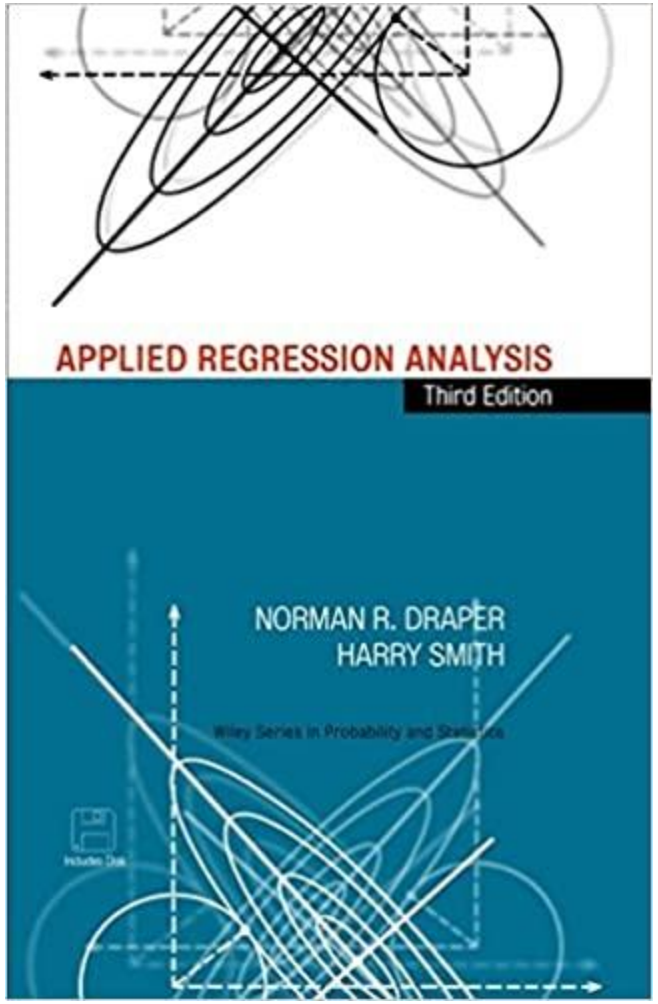
Check The Assumption

Interpret the result

* use α 5%

Purwadhika
Startup and Coding School

References



Purwadhika
Startup and Coding School

References

https://www.statsmodels.org/stable/regression.html

https://www.kaggle.com/ranjeetjain3/seaborn-tips-dataset

https://www.the-modeling-agency.com/crisp-dm.pdf

https://scikit-learn.org/stable/

Purwadhika
Startup and Coding School