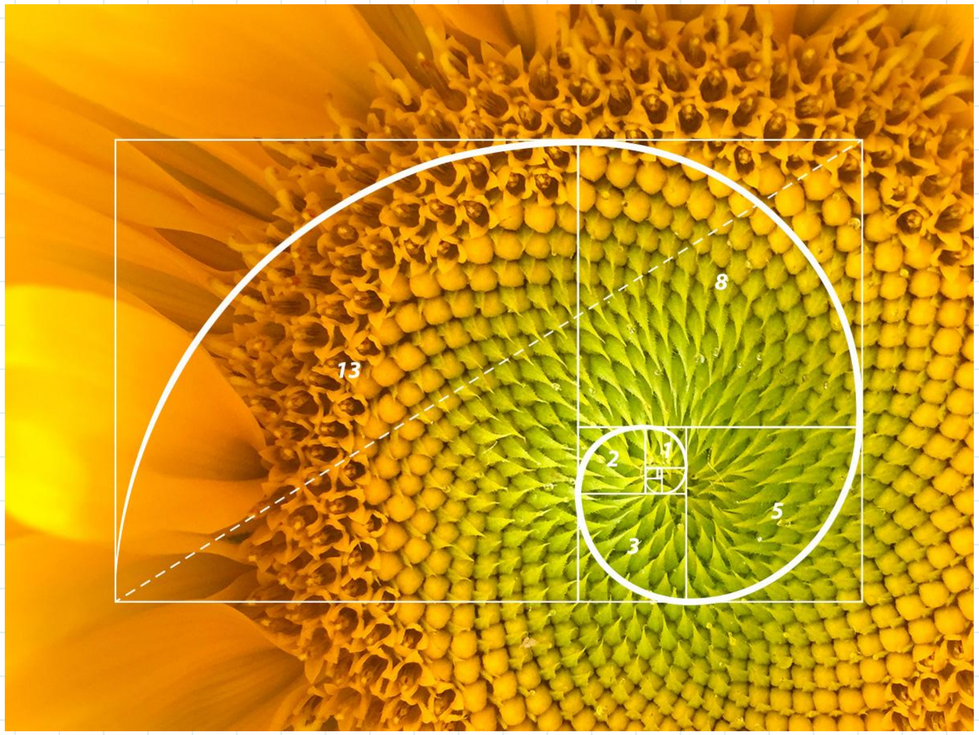
TM-03

MACHINE LEARNING

Analisis Regresi Linear

Dr. Sirojul Munir, S.Si., M.Kom.
rojulman@nurulfikri.ac.id

ARTIFICIAL INTELLIGENCE – INFORMATICS STTNF

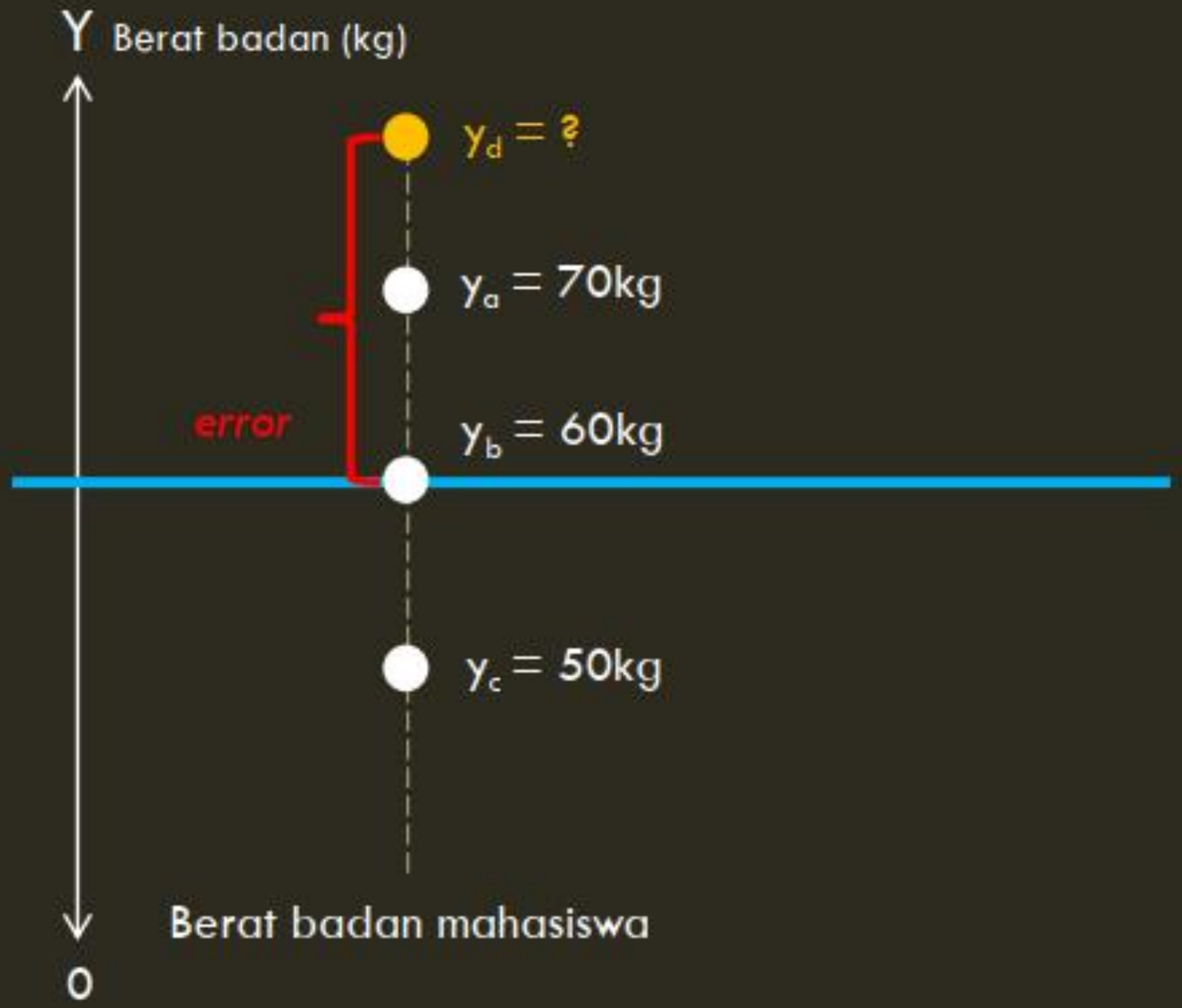


Daur ulang Project Data Science

Studi Kasus – Prediksi Berat Badan

- Misalnya Anda mengetahui rata-rata berat badan siswa di sebuah kelas dan Anda diminta memprediksi berat badan dari salah satu siswa dalam kelas tersebut yang belum pernah Anda temui, dan Anda hanya ketahui namanya saja, maka bagaimana cara Anda memberikan dugaan yang paling tepat?
- Anda dapat saja menebak dan menduga! dan ketepatan dugaan atau prediksi tersebut hanya tergantung pada keberuntungan Anda.
- Namun, dugaan yang lebih baik dan cukup mudah dihitung adalah menggunakan rata-rata berat badan mahasiswa

Prediksi menggunakan rata-rata



$\bar{y}$ = rata-rata berat badan

$$\bar{y} = \frac{y_1 + y_2 + y_3}{3}$$

$$\bar{y} = \frac{70+60+50}{3} = 60$$

Berapakah Y_d?
Kita dapat menduga nilai Y_d menggunakan rata-rata berat badan mahasiswa.

$$\hat{y}_d = \bar{y} = 60$$

Selisih antara nilai dugaan dengan nilai asli disebut error.

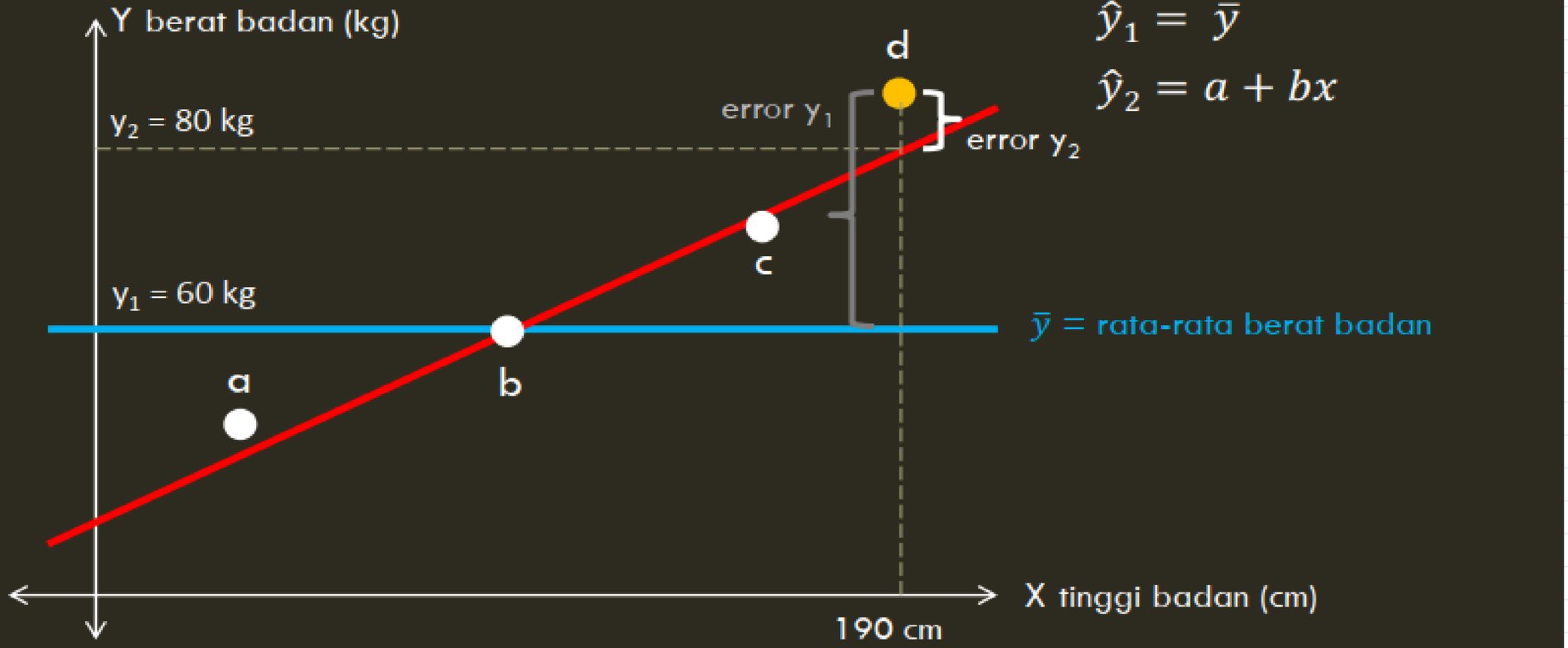
Studi Kasus – Apa itu Regresi Linear ?

- Namun, bagaimana jika Anda sudah pernah bertemu dengan mahasiswa tersebut dan melihat bahwa tinggi badannya sekitar 170 cm. Menggunakan informasi tambahan tersebut apakah Anda dapat memberikan dugaan yang lebih tepat?

- Secara fisik, tentu semakin tinggi badan maka semakin berat juga badannya. Atau dengan kata lain **tinggi ba**dan **berkorelasi** dengan **berat badan**. Dengan menggunakan data tinggi badan seseorang kita dapat membuat prediksi yang lebih akurat tentang beratbadan.

- Konsep prediksi inilah yang disebut dengan **regresi linear**.

Prediksi menggunakan Regresi



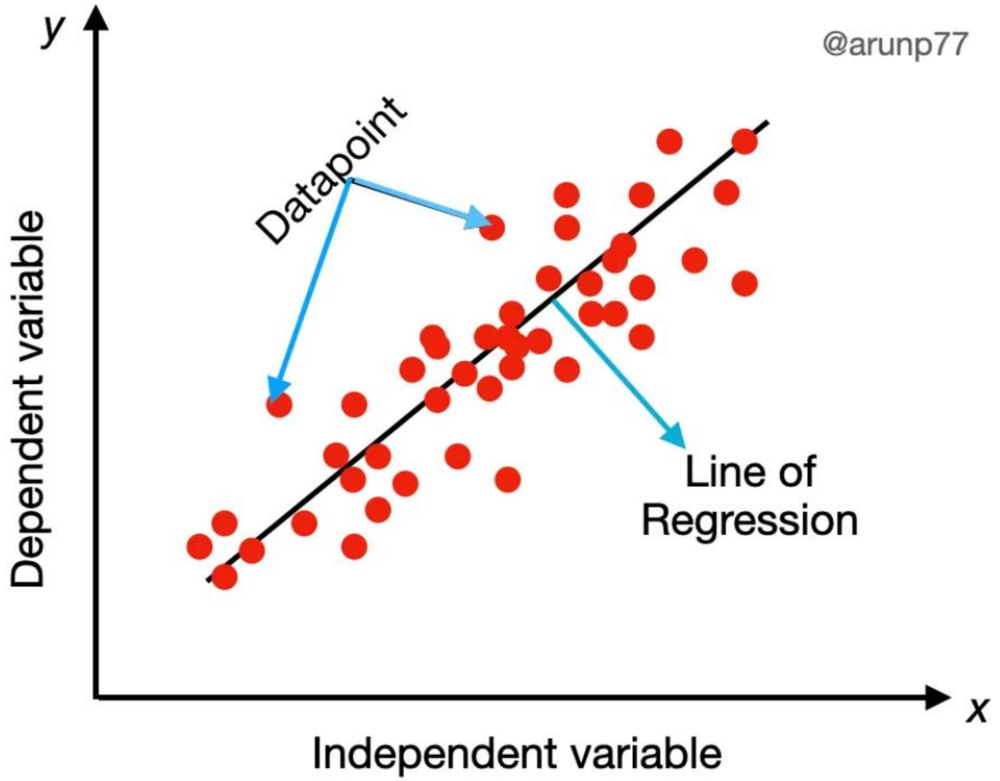
Prediksi menggunakan regresi linear (error y_2) menghasilkan total error lebih kecil dibandingkan total error dari dugaan rata-rata.

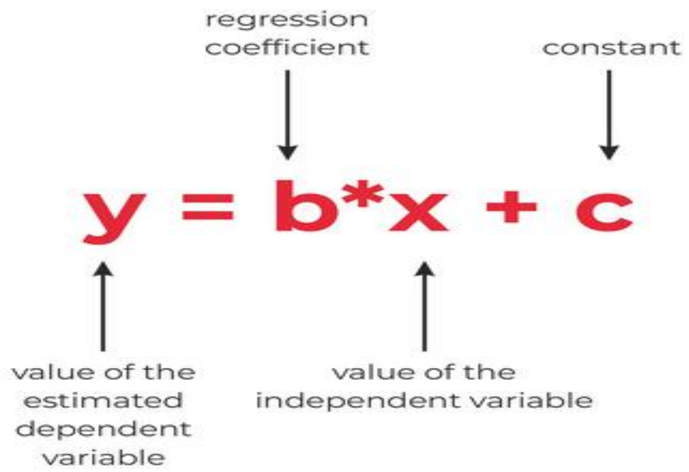
Algoritma Regresi

- Regresi adalah cara memprediksi sebuah variabel yang belum diketahui nilainya menggunakan satu atau lebih variabel lain yang sudah diketahui nilainya

- Algoritma regresi **memodelkan hubungan** antara satu atau lebih **variabel bebas** (independen) dengan **variabel target** (dependen) yang bersifat kontinu.

- **Tujuan:** memprediksi nilai variabel dependen (y) berdasarkan variabel independen (x)





Jenis Analisis Regresi

- Analisis regresi sederhana: analisis dilakukan untuk satu variabel dependen (y) terhadap satu variabel independen (x).
- Analisis multiple regresi: analisis dilakukan untuk satu variabel dependen (y) terhadap beberapa variabel independen ($x_0, x_1, x_2 \ldots$).

- Berapa jumlah variabel independen?
  - 1 : Simple regression
  - >1 : Multiple regression
- Bagaimana bentuk garis regresi?
  - Linear : Linear regression
  - Nonlinear : Nonlinear regression
- Apa jenis data variable dependen?
  - Kontinyu : Simple & Multiple regression
  - Binomial : Logistic regression

Jenis Regresi

- Linear Regression

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

- Multiple Linear Regression:

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

- Polynomial Linear Regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \varepsilon_i$$

Model Regresi Linear Sederhana

❖ Analisis regresi sederhana: analisis dilakukan untuk satu variabel dependen (y) terhadap satu variabel independen (x).

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

Where:

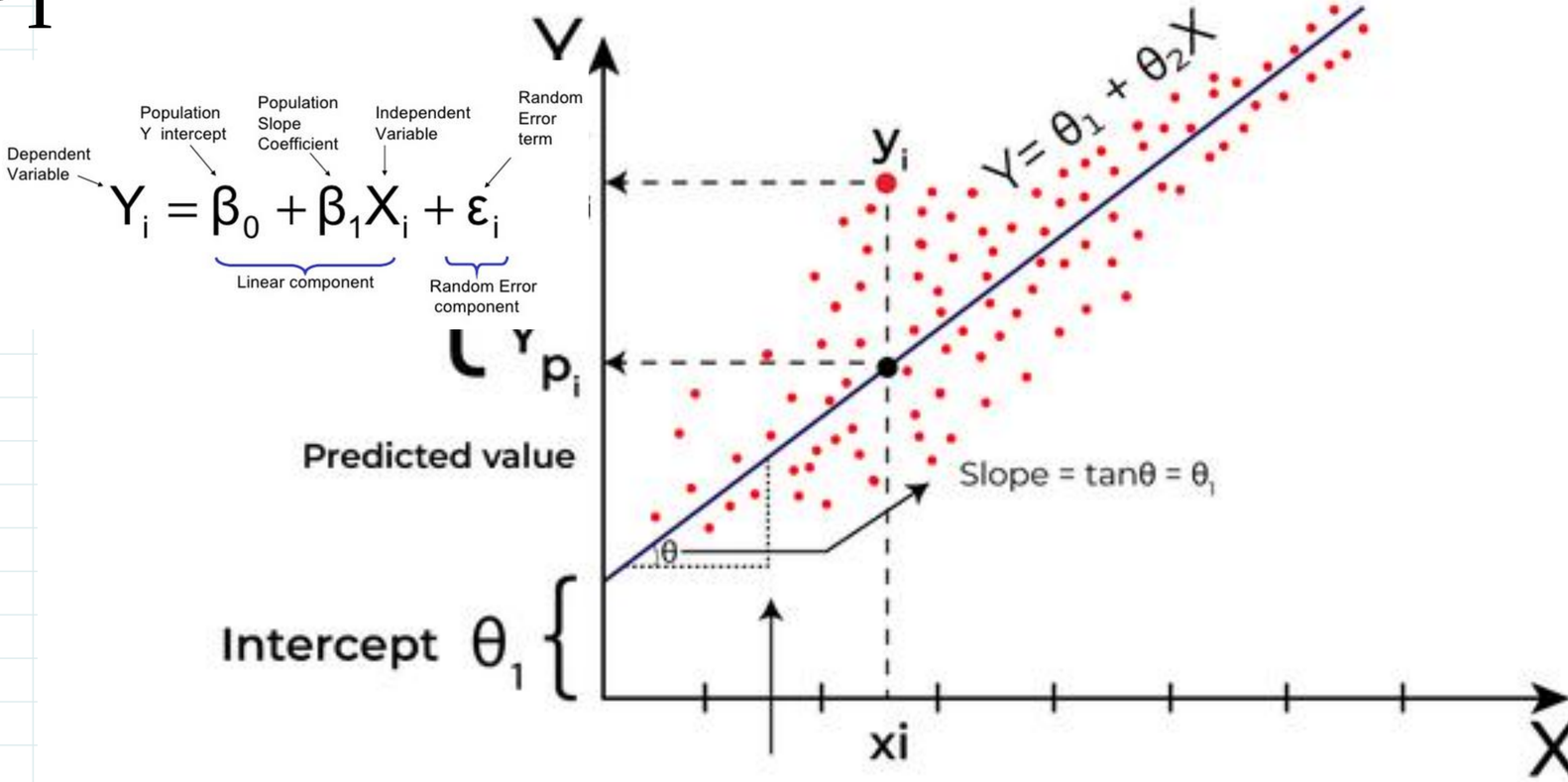
y = dependent variable

x = independet variable

β_0= y-intercept

β_1= slope of the line

ε = error variable



Model Regresi Linear Sederhana

❖ Analisis regresi sederhana: analisis dilakukan untuk satu variabel dependen (y) terhadap satu variabel independen (x).

$$Y = b_0 + bX_1 + e$$

$$\hat{Y} = b_0 + bX_1$$

Y = Berat badan actual (kg)

Y hat = berat badan yang diprediksi (kg)

X_1 = Tinggi badan (cm)

b_0 = konstanta

e = error

Model Multiple Regresi Linear

- Analisis multiple regresi: analisis dilakukan untuk satu variabel dependen (y) terhadap beberapa variabel independen ($x_0, x_1, x_2 \ldots$).

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Y = Konsumsi rumah tangga (rupiah per bulan)

Y hat = Konsumsi rumah tangga yang diprediksi (rupiah per bulan)

X_1 = Pendapatan rumah tangga (rupiah per bulan)

X_2 = Jumlah anggota keluarga (orang)

X_3 = Lokasi tempat tinggal (kota atau desa)

b_n = konstanta

$$\text{Variate (Y hat)} = X_1b_1 + X_2b_2 + \ldots + X_nb_n$$

- Nilai variat (Y hat) akan dihitung untuk setiap respon.
- Nilai Y hat adalah kombinasi linear dari seluruh gabungan variable yang menghasilkan prediksi paling baik.

Tujuan Analisis Regresi

Tiga pertimbangan utama dalam penggunaan analisis regresi:

1. Kecocokan dengan masalah yang diteliti (prediction atau explanation)
2. Penentuan *statistical relationship* (*functional relationship vs statistical relationship*)
3. Pemilihan variable dependen dan independen
   - Pastikan ada teori yang mendukung pemilihan variable.
   - Adanya *measurement error* pada variable, terutama pada variable dependen. Bisa diatasi dengan *summated scales* atau SEM
   - *Specification error* : *inclusion of irrelevant variables or exclusion of relevant variables.*

Tujuan Analisis Regresi

Analisis regresi dapat digunakan untuk:

- **Prediksi** (prediction): memperkirakan nilai sebuah variabel dependen dari beberapa variabel independen.
- **Penjelasan** (explanation): menjelaskan keterkaitan sebuah variabel dependen dengan beberapa variabel independen melalui koefisien regresi (besarnya, tandanya, signifikansinya) dan berusaha untuk mengembangkan teori mengenai efek dari variable independen terhadap variable dependen
- Regresi dapat digunakan untuk salah satu atau kedua tujuan di atas.

Rancangan Analisis Regresi (1)

❖ Beberapa pertimbangan dalam menggunakan analisis regresi

- Ukuran sampel.

  Jumlah sampel yang tersedia mempengaruhi statistical power, practical significance, dan statistical significance yang dapat dihasilkan model.

- Karakteristik hubungan variable dependen dan independen.

  Regresi linear hanya memodelkan hubungan linear. Efek non linear dapat dimodelkan dengan komponen tambahan.

- Sifat dari variable independen.

  Analisis regresi dapat mengakomodasi variable independen yang sifatnya fixed maupun yang memiliki sifat random.

Rancangan Analisis Regresi (2)

❖ Beberapa pertimbangan dalam menggunakan analisis regresi

- Analisis regresi menganalisis korelasi antara variable independen dengan variable dependen.
- Korelasi adalah seberapa besar variasi sebuah variable dapat menjelaskan variasi variable yang lain.
- Korelasi yang tinggi tidak selalu berarti kausalitas
- Contoh: Berat Badan berkorelasi tinggi dengan Tinggi Badan, namun tidak berarti berat menyebabkan tinggi.
- Untuk menganalisis kausalitas, gunakan Structural Equation Modeling (SEM)..

Rancangan Analisis Regresi

❖ Jumlah sampel sangat mempengaruhi power dan generalizability dari hasil regresi.

TABLE 5 Minimum R^2 That Can Be Found Statistically Significant with a Power of .80 for Varying Numbers of Independent Variables and Sample Sizes

	Significance Level (α) = .01 No. of Independent Variables				*Significance Level (α) = .05 No. of Independent Variables*			
Sample Size	**2**	**5**	**10**	**20**	**2**	**5**	**10**	**20**
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1,000	1	2	2	3	1	1	2	2

Note: Values represent percentage of variance explained.
NA = not applicable.

Case Study:

Dummy variable

- Sebuah analisis ingin memprediksi IPK mahasiswa (Y) dari jumlah jam belajar (X_1), jam menggunakan medsos (X_2), jam bermain game (X_3), dan gender (X_4).
- Gender adalah variable non metric dengan dua kategori yaitu pria dan wanita, yang tidak dapat digunakan langsung dalam regresi.
- Dengan demikian, untuk menggunakan variable gender dalam analisis regresi, diperlukan dummy variable, yaitu dengan mengkodekan pria = 1 dan wanita = 0 (indicator coding).
- Jika misalnya hasil regresi $Y = 1{,}5 + 2X_1 - 0{,}3X_2 - 0{,}5X_3 + 0{,}5X_4$ maka pria (indicator 1) memiliki IPK 0,5 lebih baik dibandingkan wanita (indicator 0 /reference).

Case Study:

Dummy variable

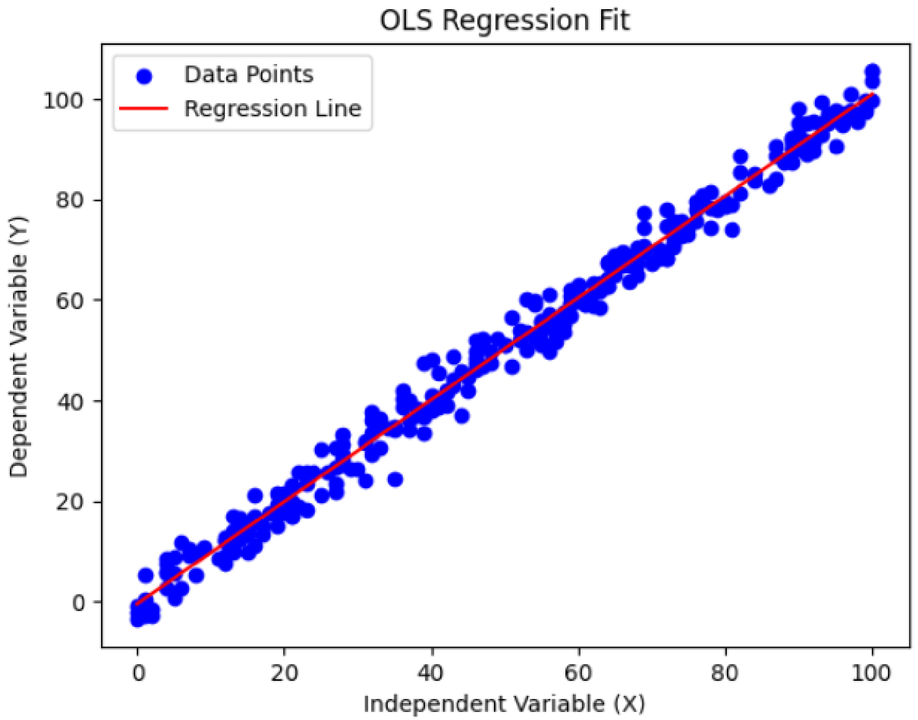
- Jika menggunakan effects coding, pria = 1 dan wanita = -1.
- Jika misalnya hasil regresi $Y = 1{,}5 + 2X_1 - 0{,}3X_2 - 0{,}5X_3 + 0{,}5X_4$ maka pria (indicator 1) memiliki IPK 0,5 lebih baik dibandingkan rata-rata seluruh mahasiswa (pria dan wanita).

Pustaka Program **Ordinary Least Squares**

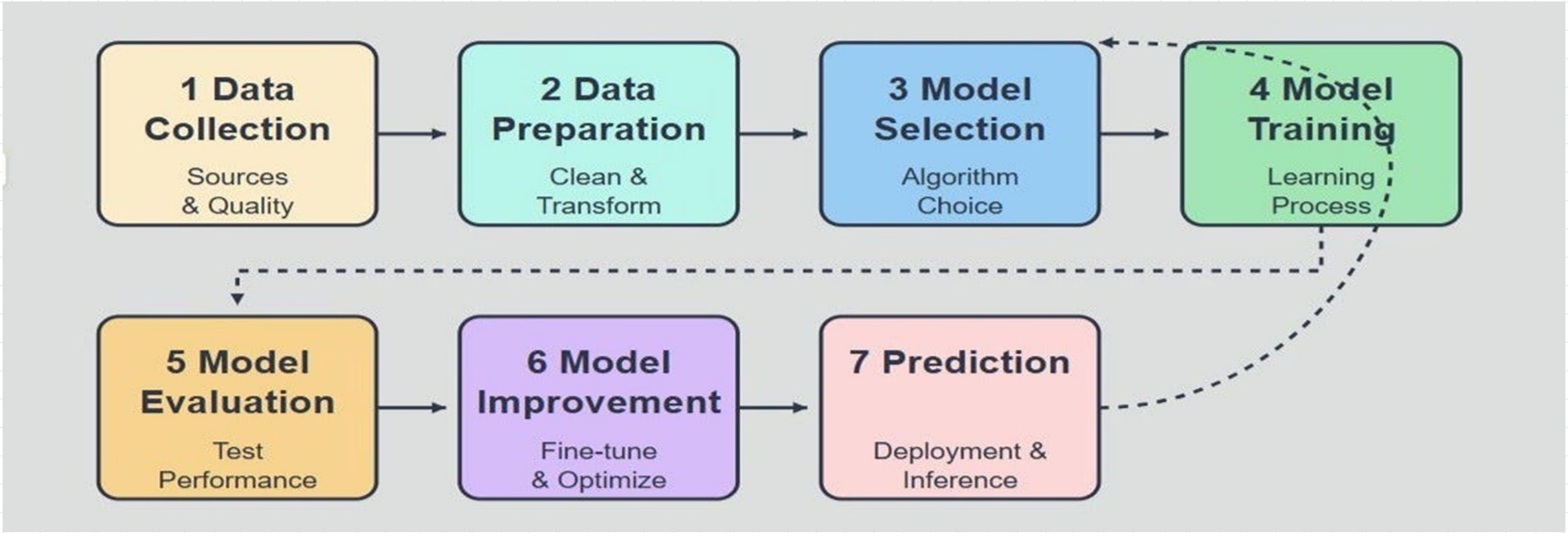
```
result = sm.OLS(y, x).fit()

print(result.summary())
```

Output :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.989
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                 2.709e+04
Date:                Tue, 08 Apr 2025   Prob (F-statistic):          1.33e-294
Time:                        05:01:34   Log-Likelihood:                -757.98
No. Observations:                 300   AIC:                             1520.
Df Residuals:                     298   BIC:                             1527.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.4618      0.360     -1.284      0.200      -1.169       0.246
x1             1.0143      0.006    164.598      0.000       1.002       1.026
==============================================================================
Omnibus:                        1.034   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.596   Jarque-Bera (JB):                0.825
Skew:                           0.117   Prob(JB):                        0.662
Kurtosis:                       3.104   Cond. No.                         120.
==============================================================================
```



Tahapan Machine Learning



Case Study: Prediksi Berat Balita

Sampel Data:

	Jenis Kelamin	Umur (bulan)	Tinggi Badan (cm)	Berat Badan (kg)	Stunting	Wasting
0	Laki-laki	19	91.6	13.3	Tall	Risk of Overweight
1	Laki-laki	20	77.7	8.5	Stunted	Underweight
2	Laki-laki	10	79.0	10.3	Normal	Risk of Overweight
3	Perempuan	2	50.3	8.3	Severely Stunted	Risk of Overweight
4	Perempuan	5	56.4	10.9	Severely Stunted	Risk of Overweight

Akan menggunakan **model OLS (Ordinary Least Squares)** untuk kasus multiple regresi Prediksi Berat badan balita berdasarkan data ~~jenis kelamin~~, umur, tinggi badan

Case Study: Korelasi antar variabel

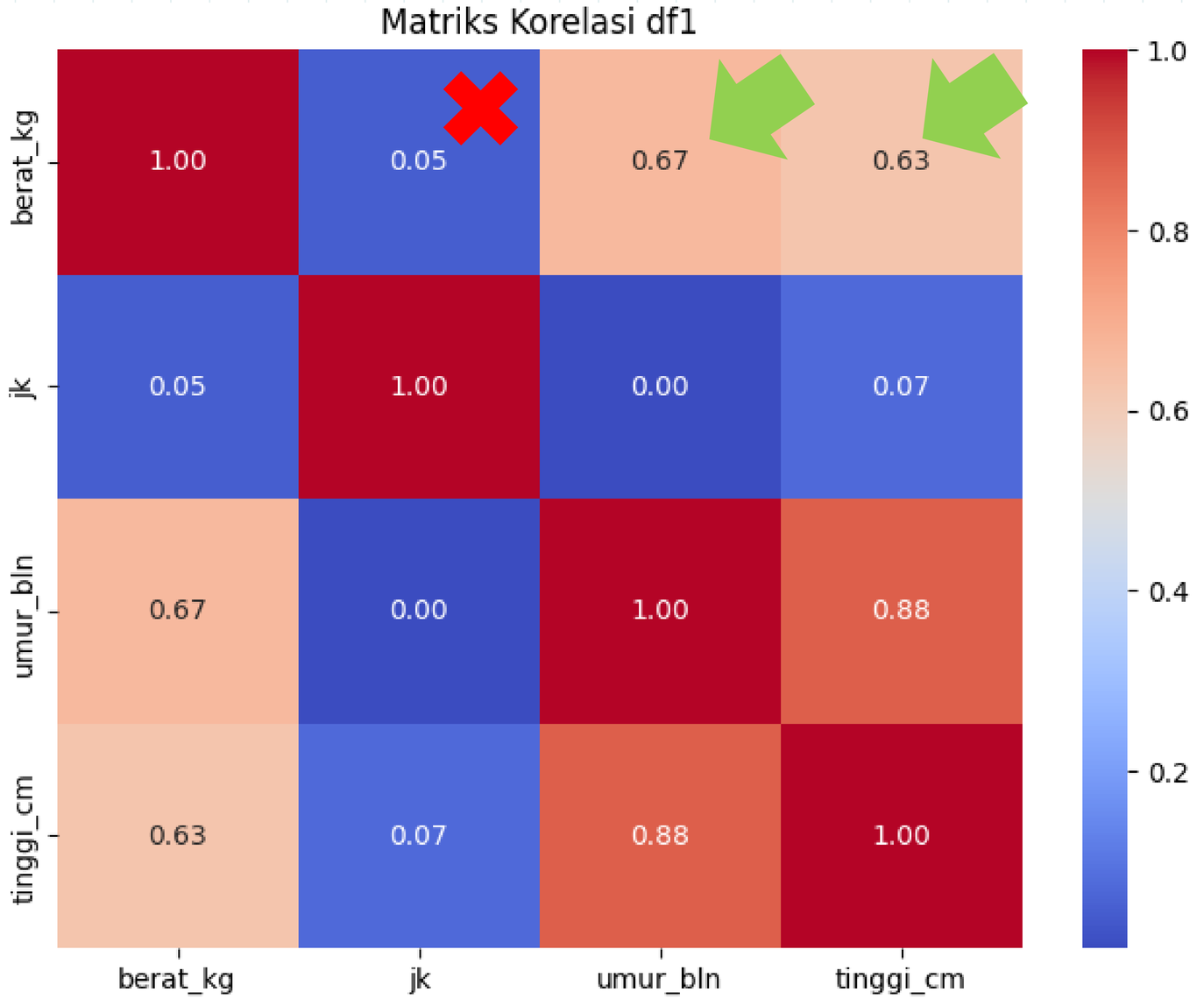
Heat Map Correlation:

Hasil Koofesian korelasi menunjukan variabel dominan yang berpengaruh dalam prediksi berat badan balita adalah:

1. Umur : 0.67 (dominan)
2. Tinggi : 0.63 (dominan)
3. Jenis kelamin : 0.05 (tidak dominan)

Karnanya model Prediksi berat Y menggunakan variabel independent

1. X1 = Umur,
2. X2 = Tinggi



Case Study: Prediksi Berat Balita

RUMUS Multiple Regresi:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

dimana:

- Y = berat badan (kg)
- X_1 = umur (bulan)
- X_2 = tinggi badan (cm)
- β_0 = intersep
- β_1, β_2 = koefisien regresi
- ε = error residual

Case Study: Prediksi Berat Balita

- const: 2.5456
- Koefisien **umur_bln** = **0.2297**
- Koefisien **tinggi_cm** = **0.0542**
- **R2 = 0.450 / 45.0%**

Hasil Model OLS – Multiple Regresi:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                berat_kg   R-squared:                       0.450
Model:                            OLS   Adj. R-squared:                  0.450
Method:                 Least Squares   F-statistic:                 3.272e+04
Date:                Sun, 05 Oct 2025   Prob (F-statistic):               0.00
Time:                        20:20:06   Log-Likelihood:            -1.8505e+05
No. Observations:               80000   AIC:                         3.701e+05
Df Residuals:                   79997   BIC:                         3.701e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5456      0.091     28.039      0.000       2.368       2.724
umur_bln       0.2297      0.002     92.330      0.000       0.225       0.235
tinggi_cm      0.0542      0.002     34.359      0.000       0.051       0.057
==============================================================================
Omnibus:                    16501.255   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000
Skew:                           0.015
Kurtosis:                       2.020
==============================================================================
```

$$\text{Berat (kg)} = 2.5456 + 0.2297 \times \text{Umur (bulan)} + 0.0542 \times \text{Tinggi (cm)}$$

1. Model cukup **baik menjelaskan 45% variasi berat badan** (R^2 = 0.45).
2. Kedua variabel (`umur_bln` dan `tinggi_cm`) **berpengaruh signifikan positif** terhadap berat.
3. Model **valid secara statistik**, karena F-test dan p-value menunjukkan signifikansi.

Tugas Praktikum Mandiri

1. Buat model prediksi dari kasus dataset berikut ini:
   https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset

[1]:
```
import pandas as pd

# Read the CSV file with a comma delimiter
df = pd.read_csv('../data/day.csv', sep=',')

# cetak header data (5 baris data) dari file
df.head()
```

[1]:

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

dengan variable dependen (Y) kolom **cnt**, tentukan variabel independent (x) dari kolom2 yang tersedia !!!

Referensi:

1. Slide presentasi Linear Regression Analysis, Harapan Bangsa Center For Data Science. 2018
2. https://www.geeksforgeeks.org/data-science/ordinary-least-squares-ols-using-statsmodels/

Terima Kasih

http://youtube.com/@rojulman