Bab 7
Linear Regression

A. KOMPETENSI DASAR

- Memahami konsep *Regression*.
- Memahami eksperimen regresi menggunakan Linear Regression
- Memahami membuat model Linear Regression

B. ALOKASI WAKTU

4 js (4x50 menit)

C. PETUNJUK

- Awali setiap aktivitas dengan do'a, semoga berkah dan mendapat kemudahan.
- Pahami Tujuan, dasar teori, dan latihan-latihan praktikum dengan baik dan benar.
- Kerjakan tugas-tugas dengan baik, sabar, dan jujur.
- Tanyakan kepada teman lalu asisten/dosen apabila ada hal-hal yang kurang jelas.

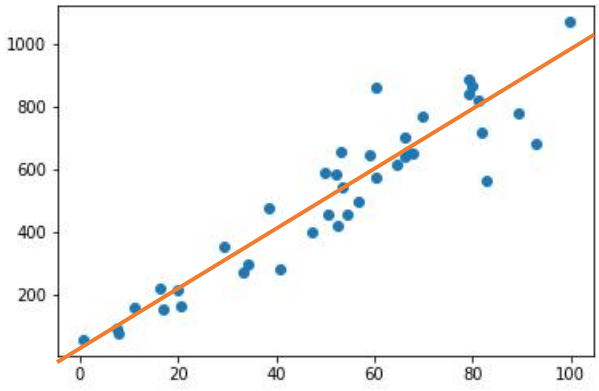
D. DASAR TEORI LINEAR REGRESSION

Istilah "linearitas" dalam aljabar mengacu pada hubungan linear antara dua atau lebih variabel. Jika kita menggambar hubungan ini dalam ruang dua dimensi (antara dua variabel), kita mendapatkan garis lurus. **Regresi linier** (*Linear Regression: LinReg*) melakukan tugas untuk memprediksi nilai variabel dependen (y) berdasarkan variabel independen (fitur atau atribut) yang diberikan (x). Jadi, teknik regresi ini menemukan hubungan linier antara x (input) dan y (output). Oleh karena itu, namanya adalah Regresi Linier. Jika kita memplot variabel independen (x) pada sumbu x dan variabel dependen (y) pada sumbu y (semua titik pada plot), regresi linier memberi kita **garis lurus** yang paling cocok dengan titik data, seperti yang ditunjukkan pada Gambar 1.

Rumus dari Gambar 1 adalah:

$$y = mx + b$$

Di mana b adalah *intercept* dan *m* adalah kemiringan garis (*slope*). Jadi pada dasarnya, algoritma regresi linier memberi kita nilai paling optimal untuk intersep dan kemiringan (dalam dua dimensi). Variabel y dan x tetap sama, karena mereka adalah fitur data dan tidak dapat diubah. Nilai-nilai yang dapat kita kontrol adalah *intercept* (b) dan *slope* (m). Terdapat kemungkinan beberapa garis lurus yang dibentuk berdasarkan pada nilai intersep dan kemiringannya. Pada dasarnya apa yang dilakukan oleh algoritma regresi linier adalah sesuai dengan banyak garis pada titik data dan mengembalikan garis yang menghasilkan kesalahan paling sedikit.



Gambar 1. Dataset Regression

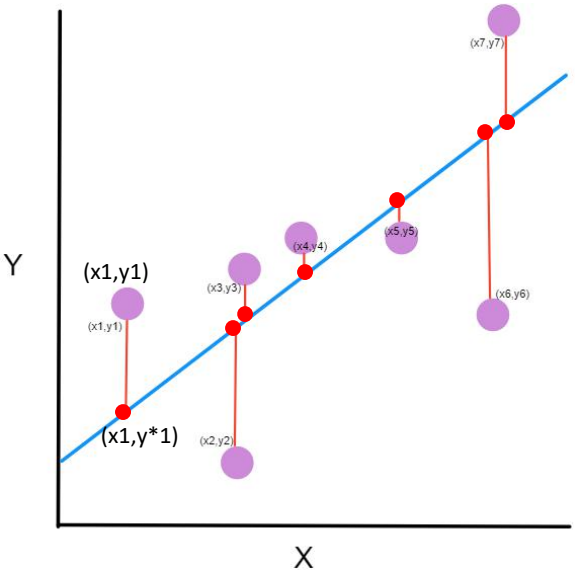
Konsep yang sama ini dapat diperluas pada kasus di mana terdapat lebih dari dua variabel. Ini disebut regresi linier berganda (*multiple linear regression*). Misalnya, pertimbangkan skenario di mana Anda harus memprediksi harga rumah berdasarkan luasnya, jumlah kamar tidur, pendapatan rata-rata orang di daerah itu, usia rumah, dan sebagainya. Dalam hal ini, variabel dependen (variabel target) bergantung pada beberapa variabel independen. Model regresi yang melibatkan banyak variabel dapat direpresentasikan sebagai:

y = b0 + m1b1 + m2b2 + m3b3 + … … mnbn

Ini adalah persamaan *hyperplane*. Ingat, model regresi linier dalam dua dimensi adalah garis lurus; dalam tiga dimensi itu adalah planar (*plane*), dan lebih dari tiga dimensi, *hyperplane*.

Dalam proses regresi linear, metrik evaluasi yang digunakan adalah Mean Squared Error (MSE). MSE dapat didefinisikan sebagai jarak absolut sebuah titik terhadap garis regresi yang dibentuk. Pada umumnya pembentukan model regresi linear ini melalui tahapan berulang atau iteratif. Sehingga target pemodelan regresi adalah total MSE terkecil dari iterasi regresi linear tersebut.

MSE = |y* - y|



Gambar 2. Ilustrasi menghitung MSE

E. LATIHAN

a) Import library yang dibutuhkan

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

b) Khusus untuk latihan membuat generator dataset.

```
def generate_dataset_simple(beta, n, std_dev):
    x = np.random.random_sample(n) * 100
    e = np.random.randn(n) * std_dev
    y = x * beta + e
    # We need to reshape x to be a 2-d matrix with n rows and 1 column
    # This is so that it can take a generalized form that can be expanded
    # to multiple predictors for the `LinearRegression` model
    return x.reshape(n, 1), y
```

c) BUat dataset dengan memanggil fungsi di atas.

```
x, y = generate_dataset_simple(10, 50, 100)

# Take the first 40 samples to train, and the last 10 to test
x_train = x[:-10]
y_train = y[:-10]

x_test = x[-10:]
y_test = y[-10:]
```

d) Plot dataset

```
plt.scatter(x_train,y_train)
plt.show()
```

e) Gunakan Library Linear Regression dari Scikit Learn untuk memodelkan garis regresi nya. Dataset telah dibagi menjadi training dan test set.

```
# Import, and create an instance of a simple least squares regression model
from sklearn import linear_model
model = linear_model.LinearRegression()

# Train the model using the training data that we created
model.fit(x_train, y_train)
print('Coefficients: \n', model.coef_)

# We then use the model to make predictions based on the test values of x
y_pred = model.predict(x_test)

# Now, we can calculate
print("Mean squared error: %.2f"
      % mean_squared_error(y_test, y_pred))
print('Variance score: %.2f' % r2_score(y_test, y_pred))
```

f) Plotting garis regresi pada training set dan test set

```
plt.scatter(x_train, y_train)
plt.plot(x_test, y_pred, color='red')
x_actual = np.array([0, 100])
y_actual = x_actual*beta
plt.plot(x_actual, y_actual, color='green')
plt.show()
```

F. TUGAS

1. Buatlah atau carilah dataset sederhana (output: harga) dengan jumlah baris antara 100-200.
2. Lakukan eksperimen (dengan 5-fold cross validation) dengan Linear Regression. Tambahkan tiap eksperimen yang anda lakukan dengan mengukur performa dari tiap model yang dihasilkan berupa: **MSE** (sklearn: mean_squared_error(yTest, yPred)) dan **Variance** (r2_score(yTest, yPred)).
3. Buatlah kesimpulan tentang Model Linear Regression berdasarkan hasil eksperimen anda.