연습문제 (Q9.1)

어느 범죄학 연구자가 인구밀도와 절도발생률 간의 관계를 연구하면서 다음의 16개 도시의 자료를 수집하였다.
X는 해당 도시의 단위면적당 인구밀도를, Y는 이전년도의 10만명당 절도범죄의 발생횟수를 조사한 것이다.
파이썬을 이용하여 회귀분석식을 구하고, 각자 회귀분석식에 대한 평가를 진행해보아라.

데이터:
X:{ 59, 49, 75, 54, 78, 56, 60, 82, 69, 83, 88, 94, 47, 65, 89, 70}
Y:{ 209, 180, 195, 192, 215, 197, 208, 189, 213, 201, 214, 212, 205, 186, 200, 204}

(1) 회귀분석 식을 구하여라 $(Y_i = \alpha + \beta X_i + \varepsilon_i)$
(2) R_squared 값을 구하여라(적합도, goodness of fit)
(3) X값이 58일때 Y값을 예측하여라.

[방법]
(1) 단일회귀분석 프로그램을 사용한다.
(2) $\alpha + \beta X_i + \varepsilon_i$ 의 각 계수를 구한다.
(3) linear_regression.score 값을 구한다.

연습문제 (Q9.2)

다중선형회귀분석 모델을 만드는 프로그램이다. 다음 프로그램을 수행해보고 아래 답하여라.
(1) **Diabetes** 데이터의 **속성**들과 데이터 개수에 대하여 설명하여라.
(2) 회귀분석을 수행하고 R-squared 값을 찾아라. 데이터를 train/test로 split하도록 프로그램을 수정하여라

(프로그램)

```
from sklearn import linear_model
from sklearn import datasets
from sklearn.metrics import mean_squared_error
import pandas as pd
diabetes_data = datasets.load_diabetes()
X = pd.DataFrame(diabetes_data.data)
y = diabetes_data.target
linear_regression = linear_model.LinearRegression()
linear_regression.fit(X = pd.DataFrame(X), y = y)
prediction = linear_regression.predict(X = pd.DataFrame(X))
print('a value = ', linear_regression.intercept_)
print('b balue =', linear_regression.coef_)
residuals = y-prediction
SSE = (residuals**2).sum(); SST = ((y-y.mean())**2).sum()
R_squared = 1 - (SSE/SST)
print('R_squared = ', R_squared)
print('score = ', linear_regression.score(X = pd.DataFrame(X), y = y))
print('Mean_Squared_Error = ', mean_squared_error(prediction, y))
print('RMSE = ', mean_squared_error(prediction, y)**0.5)
```