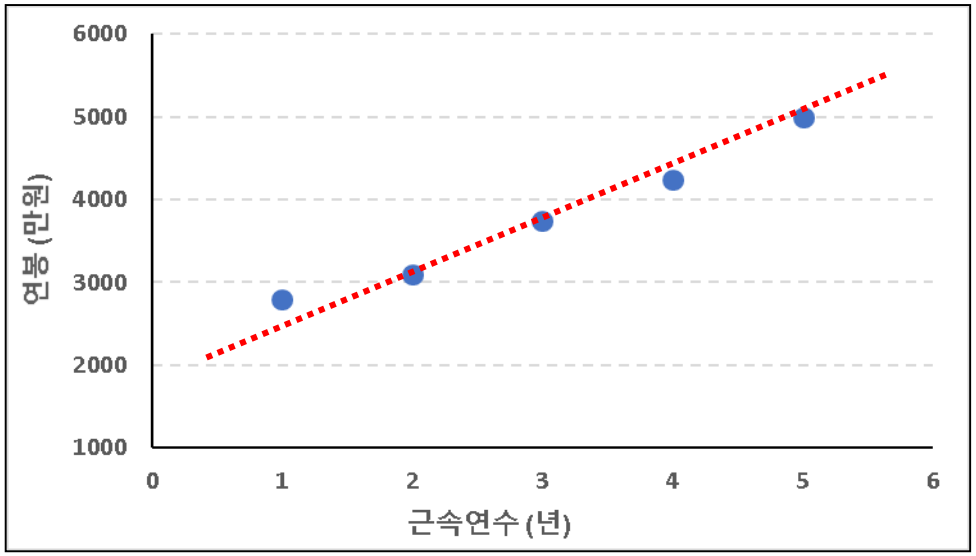
회귀 분석

회귀 분석

- 회귀 (Regression)
  - 데이터의 값은 평균과 같은 기존의 경향으로 돌아가려는 경향이 있다는 것
  - 여러 변수들 간의 상관 관계를 파악하여, 어떤 특정 변수의 값을 다른 변수들의 값을 이용하여 설명/예측하는 기법

독립변수 **종속변수** **회귀식: 연봉 = 554×근속연수 + 2116**

근속연수 (년)	연봉 (만원)
1	2800
2	3100
3	3750
4	4240
5	5000



회귀 분석

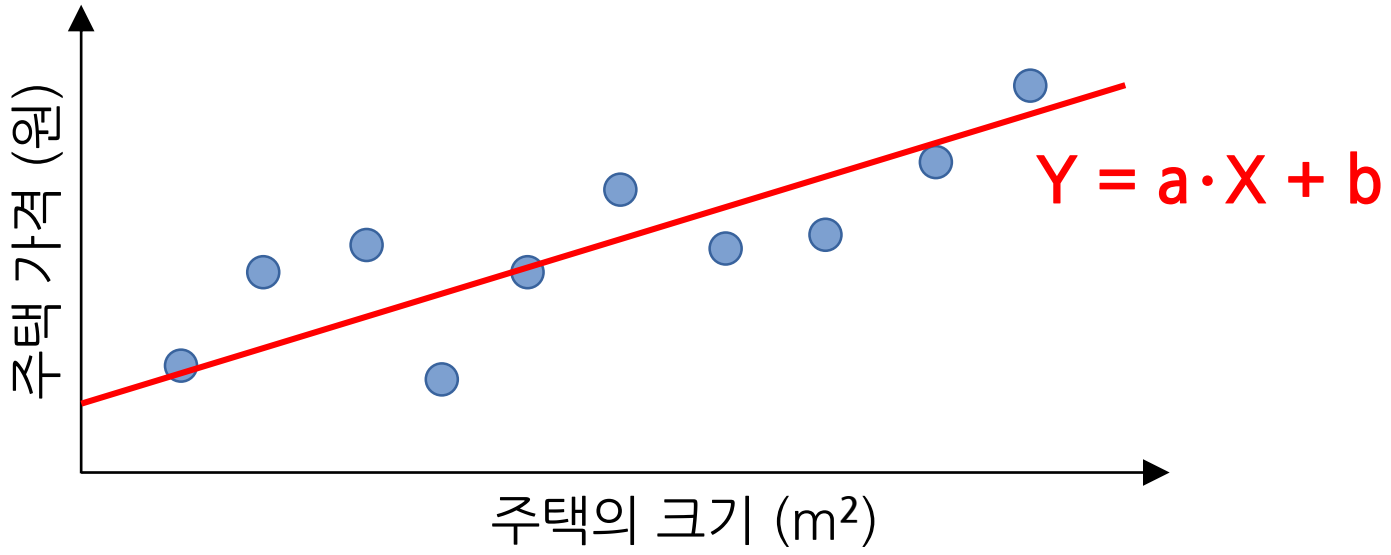
- 회귀 분석의 유형
  - 변수의 개수 및 계수의 형태에 따라 구분한다.

  - 독립변수의 개수에 따라
    - 단순 : 독립변수가 1개인 경우
    - 다중 : 독립변수가 여러 개인 경우
  - 회귀계수의 형태에 따라
    - 선형 : 계수를 선형 결합으로 표현할 수 있는 경우
    - 비선형 : 계수를 선형 결합으로 표현할 수 없는 경우

단순 선형 회귀

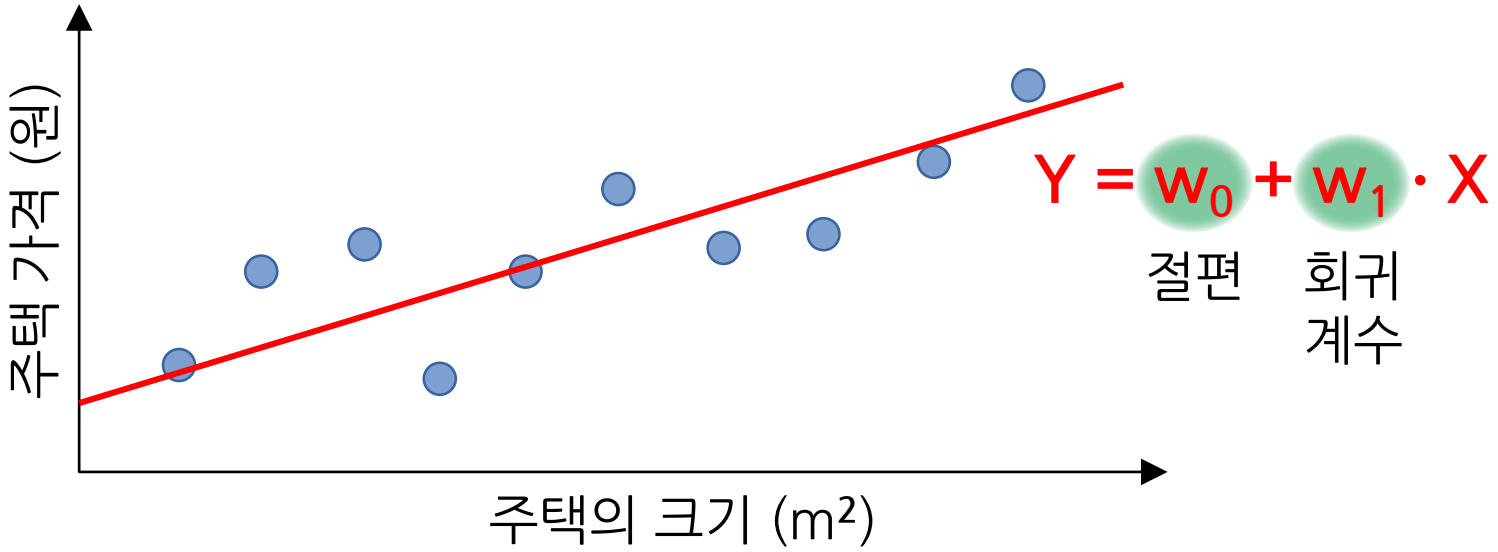
단순 선형 회귀

- 단순 선형 회귀 (Simple Linear Regression)
  - 독립변수가 1개이고 종속변수도 1개인 경우, 그들 간의 관계를 선형적으로 파악하는 회귀 방식
  - 독립변수 X와 종속변수 Y의 관계를 Y = aX + b 형태의 1차 함수식으로 표현할 수 있다.



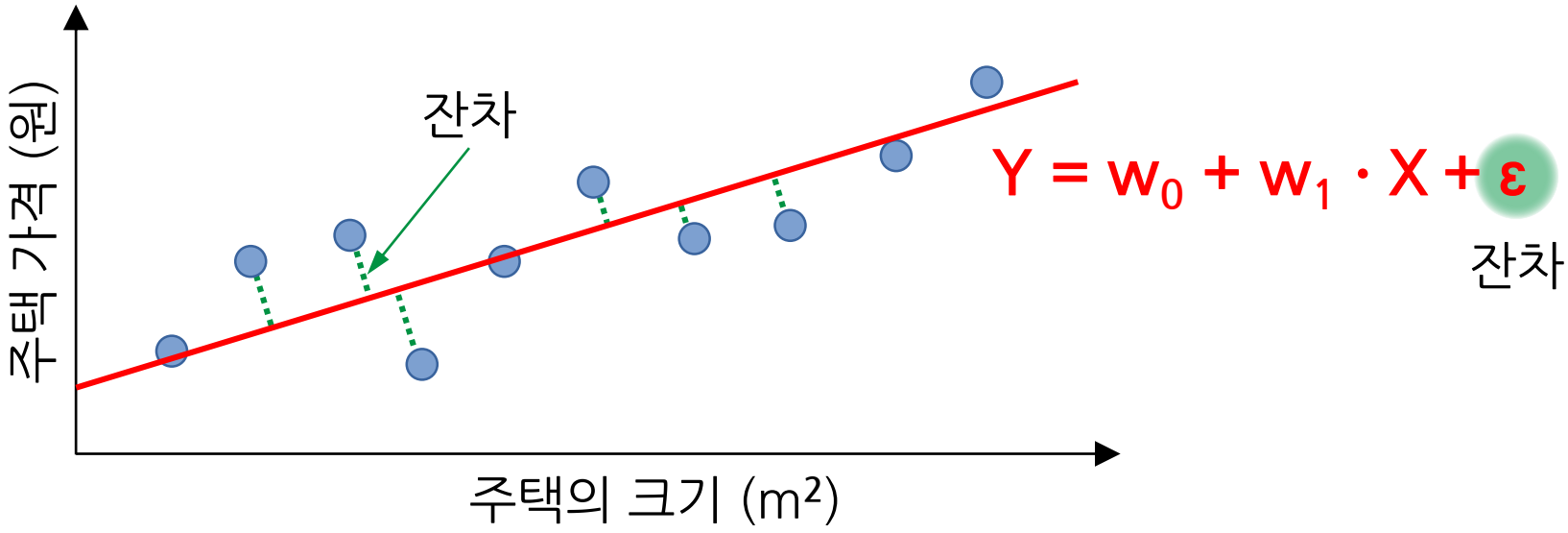
단순 선형 회귀

- 단순 선형 회귀
  - 회귀 계수 (coefficient)
    - 독립변수가 종속변수에 끼치는 영향력의 정도로서, 직선의 기울기(slope)
  - 절편 (intercept)
    - 독립변수가 0일 때의 상수 값



단순 선형 회귀

- 단순 선형 회귀
  - 잔차 (residual)
    - 실제 값과 회귀식의 차이에 따른 오류 값
    - 잔차 값이 작을수록, 구해진 회귀식이 데이터들을 더욱 잘 설명하고 있다고 볼 수 있다.



단순 선형 회귀

- 단순 선형 회귀
  - 잔차제곱합 (RSS; Residual Sum of Squares)
    - 잔차는 양수 또는 음수가 될 수 있는 값이므로 이들을 단순히 더하면 안 되고,이 값들의 제곱을 구해서 더한다.

$$RSS = \sum (y_i - (w_0 + w_1 \cdot x_i))^2$$

(이 때, x_i는 독립변수 집합 X의 원소, y_i는 종속변수 집합 Y의 원소이다.)

    - 이 때, RSS를 회귀 분석에서의 손실 함수(loss function) 또는 비용 함수(cost function)라고 한다.
  - 최적의 회귀 모형을 만든다는 것은 RSS 값이 최소가 되는 회귀 계수를 구한다는 의미이다.

회귀 분석의 평가 지표

- 회귀 분석 결과에 대한 주요 평가 지표

지표	의미	수식	대응 함수
MAE	Mean Absolute Error, 즉 실제값과 예측값의 차이의 절대값들의 평균	$\frac{1}{N}\sum \lvert y_i - \hat{y}_i \rvert$	metrics 모듈의 mean_absolute_error
MSE	Mean Squared Error, 즉 실제값과 예측값의 차이의 제곱들의 평균	$\frac{1}{N}$ RSS	metrics 모듈의 mean_squared_error
RMSE	Root of MSE, 즉 MSE의 제곱근 값	$\sqrt{MSE}$	math 또는 numpy 모듈의 sqrt
R^2	결정 계수라고 하며, 실제값의 분산 대비 예측값의 분산의 비율	$\frac{\text{예측값 분산}}{\text{실제값 분산}}$	metrics 모듈의 r2_score 또는 LinearRegression의 score

※ 이 때, $\hat{y}_i$는 실제값 y_i에 대한 예측값이다.

회귀 분석의 평가 지표

- 결정 계수 (Coefficient of Determination)
  - 회귀식이 얼마나 설명력이 있는지 (즉, 얼마나 정확한지) 나타내는 지표이다.

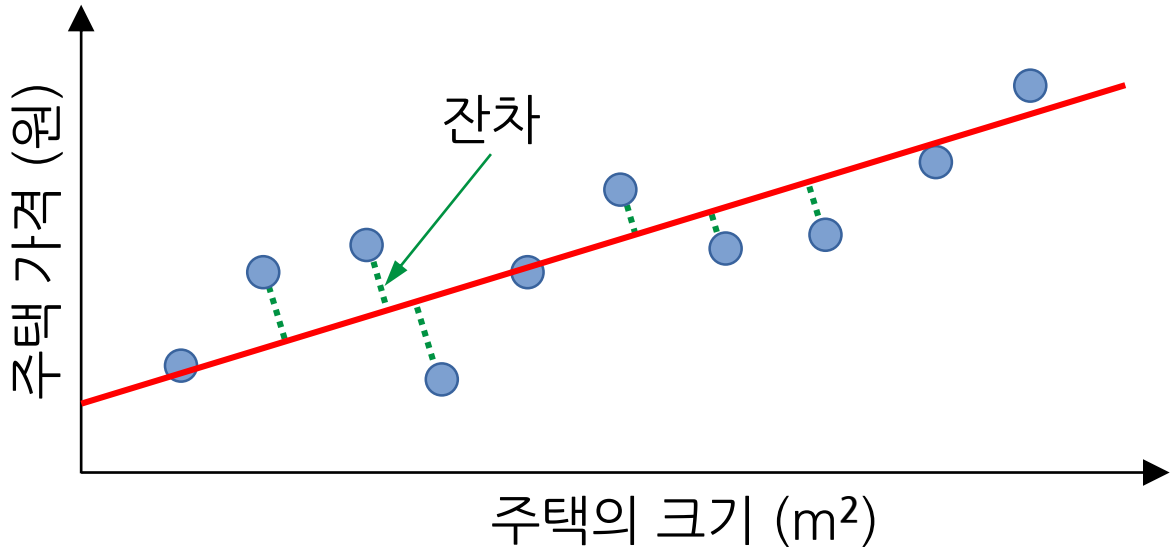
$$R^2 = \frac{\text{예측값의 분산}}{\text{실제값의 분산}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{\sum(y_i - \bar{y})^2}$$

(이 때, $\hat{y}_i$는 실제값 y_i에 대한 예측값, $\bar{y}$는 실제값들의 평균이다.)

  - 결정 계수의 값은 $0 \leq R^2 \leq 1$이며, 1에 가까울수록 설명력이 강하고 0에 가까울수록 설명력이 약하다.
  - 일반적으로 결정 계수 R^2의 값이 0.65 (65%) 이상이면 설명력이 있다고 판단한다.

단순 선형 회귀

- 최소제곱법 (OLS; Ordinary Least Squares)
  - 잔차제곱합 RSS 값이 최소화 되도록 손실 함수의 매개변수 w_0와 w_1의 값을 구한다.
  - w_0와 w_1으로 RSS 함수를 각각 편미분한 값이 0이 되는 연립 방정식의 해를 구한다.



단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행
  ① linear_model 모듈에 있는 LinearRegression을 이용하여 OLS 방법으로 선형 회귀를 수행할 수 있는 객체를 생성한다.
  - 이 때 다음과 같은 매개변수들을 추가 설정할 수 있으나, 대부분의 경우에는 필요하지 않다.
    - fit_intercept : 절편 값을 계산할 것인지의 여부를 결정한다. 기본값은 True이다.
    - normalize : 회귀를 수행하기 전에 데이터를 정규화할 것인지의 여부를 결정한다. 기본값은 False이다.

```
1  import sklearn.linear_model as lm
2
3  lr = lm.LinearRegression()
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행

  ② 선형 회귀를 수행할 객체에 대하여 **fit** 메소드를 이용하여 학습을 수행하여 회귀 모형을 추정한다.

  - 첫 번째 매개변수는 학습용 데이터의 독립변수 집합이다.
  - 두 번째 매개변수는 학습용 데이터의 종속변수 집합이다.

```
1  X_train = [[1], [2], [3], [4], [5]]
2  y_train = [2.3, 3.99, 5.15, 7.89, 8.6]
3
4  reg = lr.fit(X_train, y_train)
```

※ 독립변수의 특성이 1개 밖에 없더라도 각 값들은 리스트 또는 배열의 형태여야 한다.

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행
  ③ 실행 객체 또는 추정된 회귀 모형에 대하여 **predict** 메소드를 이용하여 예측을 수행한다.
  - 매개변수는 검증용 데이터의 독립변수 집합이다.
  - 반환 결과는 검증용 데이터에 대한 종속변수 예측값이다.

```
1  X_test = [[6], [7]]
2  y_test = [10.1, 11.9]
3
4  y_pred = reg.predict(X_test)
```

```
1  print(y_pred)
```

```
[10.536 12.186]
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행

  ④ 분석 결과를 평가한다. (MSE 및 RMSE)

  - **metrics** 모듈에 있는 **mean_squared_error** 함수를 이용하여 MSE를 구한다.

```
1  import sklearn.metrics as mt
2
3  mse = mt.mean_squared_error(y_test, y_pred)
4  print("MSE: {:.3f}".format(mse))
```

```
MSE: 0.136
```

  - MSE의 제곱근을 계산하여 RMSE를 구한다.

```
1  import numpy as np
2
3  rmse = np.sqrt(mse)
4  print("RMSE: {:.3f}".format(rmse))
```

```
RMSE: 0.369
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행

  ④ 분석 결과를 평가한다. (결정 계수 R^2)

  - **metrics** 모듈에 있는 **r2_score** 함수를 이용하여 결정 계수 R^2 값을 구한다.
  - 이 때 첫 번째 매개변수는 검증용 데이터의 종속변수 실제값이고, 두 번째 매개변수는 종속변수 예측값이다.

```
1  r2 = mt.r2_score(y_test, y_pred)
2  print("R2: {:.3f}".format(r2))
```

```
R2: 0.832
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행

  ④ 분석 결과를 평가한다. (결정 계수 R^2)

  - 또는, 실행 객체 또는 추정된 회귀 모형에 대하여 **score** 메소드를 호출하여 R^2 값을 구할 수도 있다.
  - 이 때 첫 번째 매개변수는 검증용 데이터의 독립변수이고, 두 번째 매개변수는 종속변수이다.

```
1  r2 = reg.score(X_test, y_test)
2  print("R2: {:.3f}".format(r2))
```

```
R2: 0.832
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행
  ④ 분석 결과를 평가한다. (회귀 계수 및 절편)
  - 추정된 회귀 모형의 회귀 계수 및 절편 값을 확인한다.
  - 회귀 계수는 **coef_** 속성, 절편은 **intercept_** 속성에 각각 값이 할당되어 있다.

```
1 print("회귀 계수:", reg.coef_)
2 print("절편:", reg.intercept_)
```

```
회귀 계수: [1.65]
절편: 0.6359999999999983
```

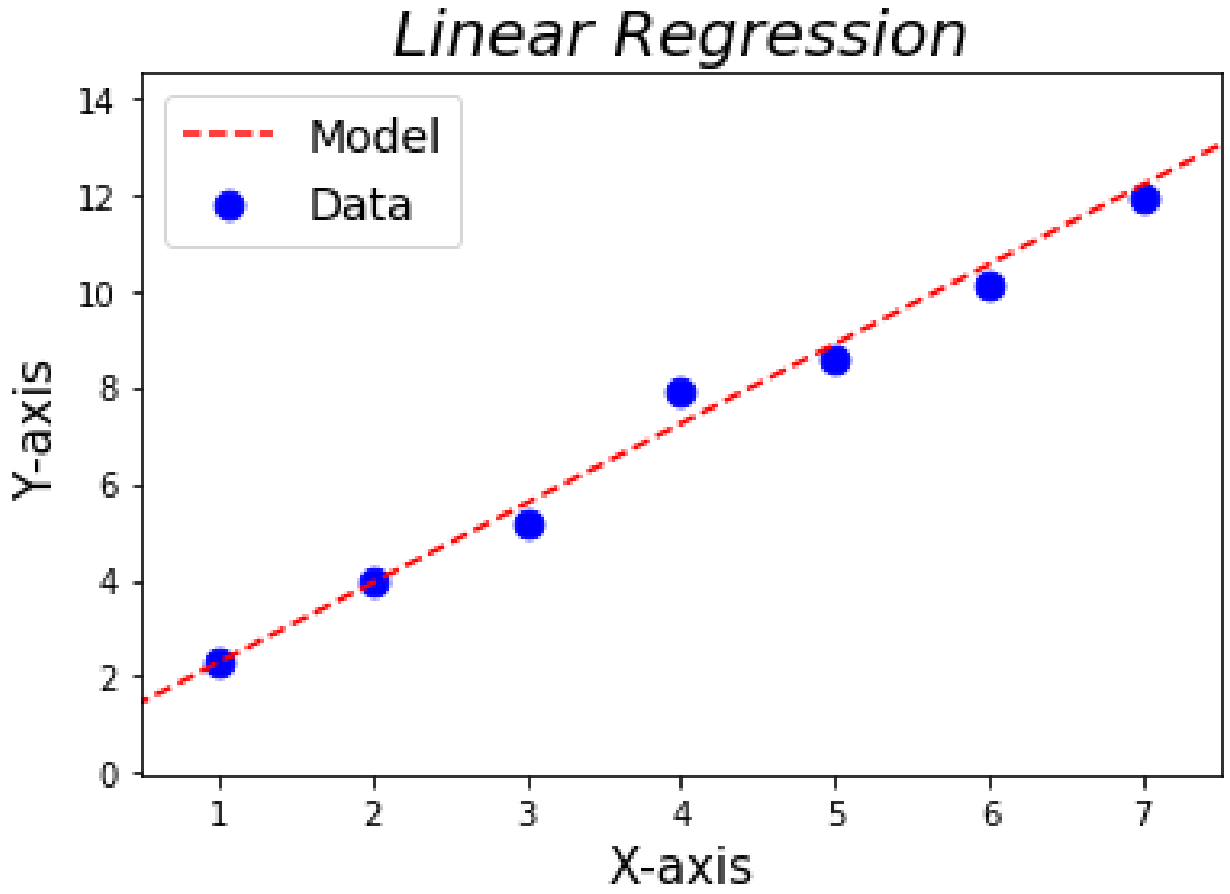
```
1 print("회귀식: y = {:.2f} X + {:.3f}".format(reg.coef_[0], ₩
2                                              reg.intercept_))
```

```
회귀식: y = 1.65 X + 0.636
```

단순 선형 회귀

- 사이킷런에서 최소제곱법으로 단순 선형 회귀 수행

  ⑤ 분석 결과를 플롯으로 표현해 본다.



단순 선형 회귀

- 스탯츠모델에서 최소제곱법으로 단순 선형 회귀 수행
  ① api 모듈에 있는 add_constant 함수를 이용하여 상수항을 추가하도록 지정한다.
  - 매개변수는 학습용 데이터의 독립변수 집합이다.
  - 반환 결과는 회귀 모형에 상수항이 추가되도록 변형된 독립변수 집합이다.

```
1  import statsmodels.api as sm
2
3  X_train = [[1], [2], [3], [4], [5]]
4  y_train = [2.3, 3.99, 5.15, 7.89, 8.6]
5
6  X_train = sm.add_constant(X_train)
7  print(X_train)
```

```
[[1. 1.]
 [1. 2.]
 [1. 3.]
 [1. 4.]
 [1. 5.]]
```

단순 선형 회귀

- 스탯츠모델에서 최소제곱법으로 단순 선형 회귀 수행

② api 모듈에 있는 OLS를 이용하여, 선형 회귀를 수행할 수 있는 객체를 생성한다.

- 첫 번째 매개변수는 학습용 데이터의 종속변수 집합이다.
- 두 번째 매개변수는 학습용 데이터의 독립변수 집합이다.

③ 선형 회귀를 수행할 객체에 대하여 fit 메소드를 이용하여 학습을 수행하여 회귀 모형을 추정한다.

- 학습용 데이터들을 이미 객체에 넣어 주었기 때문에, 매개변수로 데이터를 전달하지 않는다.

```
1  lr = sm.OLS(y_train, X_train)
2  reg = lr.fit()
```

단순 선형 회귀

- 스탯츠모델에서 최소제곱법으로 단순 선형 회귀 수행

  ④ 추정된 회귀 모형에 대하여 **predict** 메소드를 이용하여 예측을 수행한다.

  - 매개변수는 검증용 데이터의 독립변수 집합으로서, 학습 때와 마찬가지로 미리 상수항을 추가시켜야 한다.
  - 반환 결과는 검증용 데이터에 대한 종속변수 예측값이다.

```
1  X_test = [[6], [7]]
2  y_test = [10.1, 11.9]
3
4  X_test = sm.add_constant(X_test)
5  y_pred = reg.predict(X_test)
```

```
1  print(y_pred)
```

```
[10.536 12.186]
```

단순 선형 회귀

- 스탯츠모델에서 최소제곱법으로 단순 선형 회귀 수행

  ⑤ 추정된 회귀 모형에 대하여 **summary** 메소드를 이용하여 분석 결과를 평가한다. (R^2, 회귀 계수, 기타 검정 통계량 등)

```
1  print(reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.966
Method:                 Least Squares   F-statistic:                     116.2

==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6360      0.508      1.253      0.299      -0.979       2.251
x1             1.6500      0.153     10.781      0.002       1.163       2.137
```

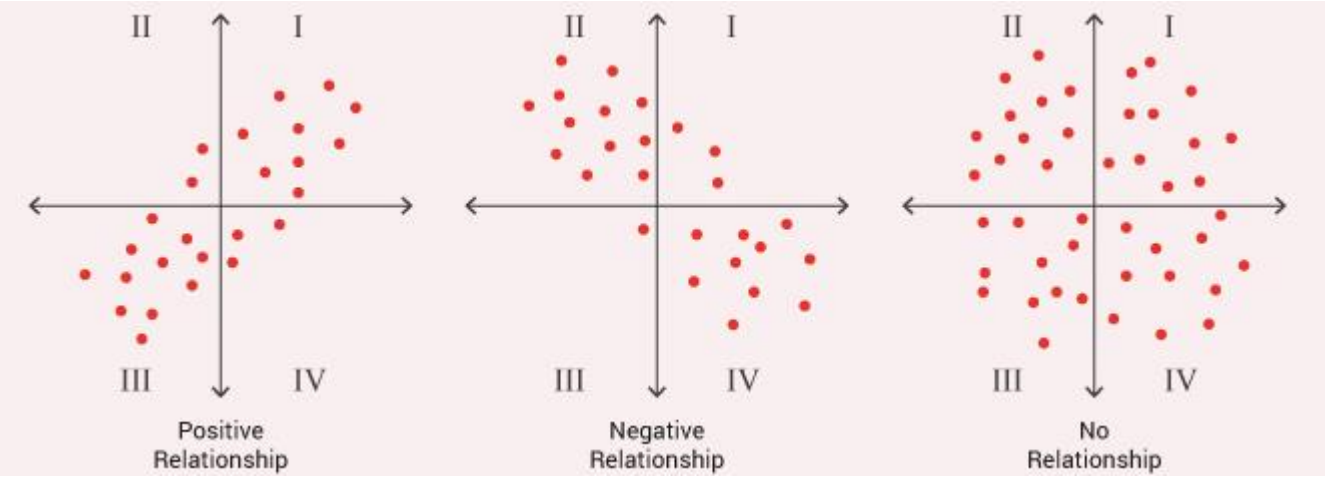
참고 : 결정 계수 R^2에 대한 이해

평가 지표에 대한 이해

- 공분산 (Covariance)
  - 2개의 변수들 간의 상관 관계를 나타낸 수치
    - 2개의 변수 X, Y에 대하여 X(또는 Y)의 값이 변화할 때 Y(또는 X)의 값이 어떻게 분포되는가를 나타낸다.

$$Cov(X, Y) = E((X-\mu)\times(Y-v))$$

(이 때, μ는 X의 평균 E(X)이고, v는 Y의 평균 E(Y)이다.)



※ 그림 출처: https://acadgild.com/blog/covariance-and-correlation

평가 지표에 대한 이해

- 상관 계수 (Correlation Coefficient)
  - 공분산을 각각의 표준편차로 나누어 정규화한 수치
    - 변수 X, Y에 대하여 각각의 크기(단위)에 영향을 받지 않도록 단위를 보정한 것이라고 볼 수 있다.

$$R = \frac{Cov(X, Y)}{\sigma_X \times \sigma_Y}$$

(이 때, σ_X는 X의 표준편차, σ_Y는 Y의 표준편차이다.)

    - 상관 계수의 값은 $-1 \leq R \leq 1$이며, 1에 가까울수록 강한 양(+) 의 상관 관계, -1에 가까울수록 강한 음(-)의 상관 관계이다. 0이면 서로 상관 관계가 없다.

평가 지표에 대한 이해

- 결정 계수 (Coefficient of Determination)
  - (단순 선형 회귀에서) 상관 계수를 제곱한 수치
    - 한 변수의 변화량이 다른 변수의 변화량으로 얼마나 설명이 될 수 있는지를 나타낸다.

$$R^2 = \frac{\text{예측값의 분산}}{\text{실제값의 분산}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{\sum(y_i - \bar{y})^2}$$

(이 때, $\hat{y}_i$는 실제값 y_i에 대한 예측값, $\bar{y}$는 실제값들의 평균이다.)

    - 결정 계수의 값은 $0 \leq R^2 \leq 1$이며, 1에 가까울수록 설명력이 강하고 0에 가까울수록 설명력이 약하다.