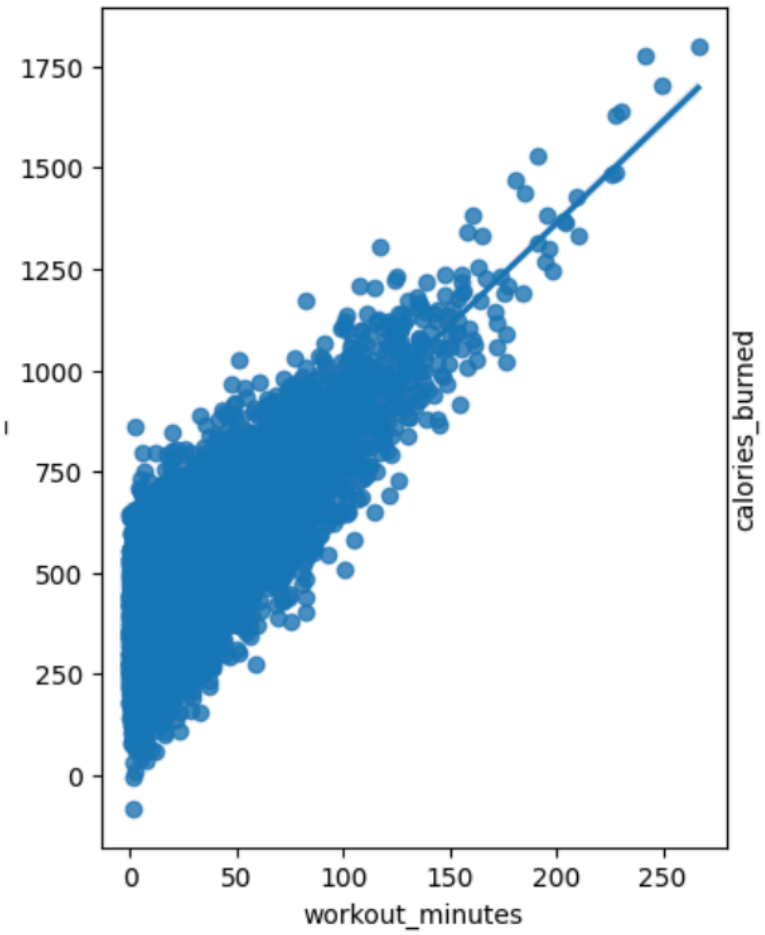
5회차

🧠 회귀(Regression)란

- 연속형의 결과값을 예측하는 기법
- 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법
    - 키 : 부모가 크면 자식도 크고, 부모가 작으면 자식도 작은 경향은 있으나 세대를 이어가며 자식이 무한히 커지거나 자신이 무한히 작아지지는 않음. 사람의 키는 평균 키로 회귀하려는 경향을 가짐.
- 1개 이상의 독립변수(X)와 종속 변수(Y) 간의 관계를 모델링
    - X와 Y 사이의 관계식을 만드는 것이다

쉽게 표현하면, 산점도에서 가장 좋은 직선을 긋는 것!



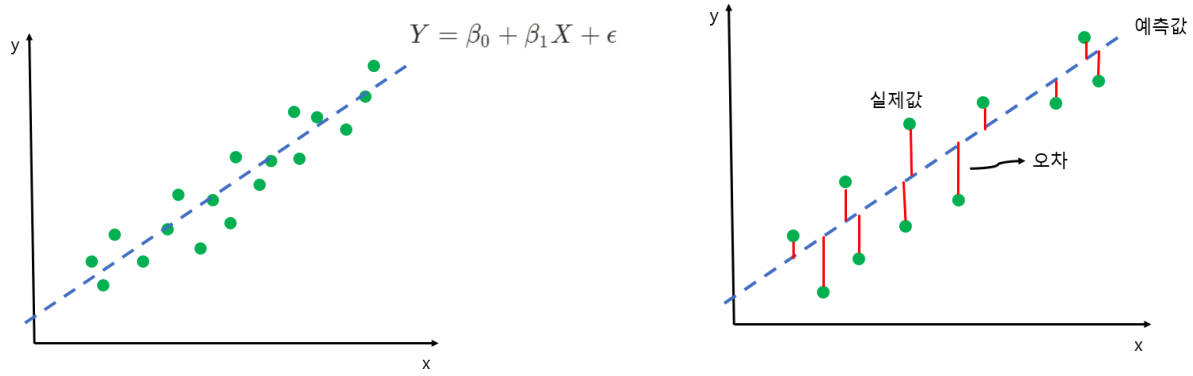
선형 회귀모델의 기본 형태

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

- Y_i: 종속변수 (예측하고자 하는 값)
- X_i: 독립변수 (설명 변수)
- β_0: 절편 (intercept)
- β_1: 기울기 (slope)
- ε_i: 오차항 (error term), 평균 0, 분산 σ^2

회귀모델의 기본원리

: 가장 실제 값에 근사한 예측값을 찾아내기 → 실제값과의 예측값 사이의 **오차를 최소화** 해야 한다. → 즉, 오차를 최소화하는 β_0, β_1(회귀계수)를 찾는다.



- 최소자승법(Ordinary Least Sqaures)
  - 미분을 통해 잔차제곱합 $\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$을 최소화하는 β_0, β_1 을 구함.

    $\hat{\beta}1 = \frac{\sum i=1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

    $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

주요 기본 개념

독립변수/종속변수

구분	영어	의미
독립변수	Independent Variable (X)	종속변수에 영향을 주는 **설명 변수** **- 분석의 수단이 되는 변수**
종속변수	Dependent Variable (Y)	우리가 **예측하려는 목표 변수** **- 분석의 대상이 되는 변수**

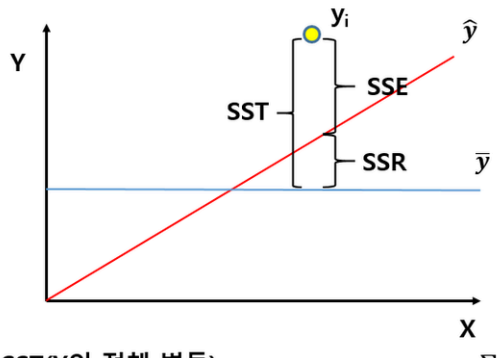
상황	독립변수(X)	종속변수(Y)
광고비와 매출 관계	광고비	매출액
공부시간과 시험점수 관계	공부시간	시험점수
근무연수와 연봉 관계	근무연수	연봉

회귀계수/ 절편

항목	설명	해석 예시
절편(β_0)	X가 0일 때 Y의 예상값	광고비가 0일 때 예상 매출액
회귀계수(β_1)	X가 1 증가할 때 Y의 변화량	광고비가 1만 원 증가할 때 매출은 1.4만 원 증가

결정계수 (R^2, R-squared)

- 모델이 데이터를 **얼마나 잘 설명하는가(설명력)** 를 나타내는 지표입니다.



SST(Y의 전체 변동) : $\sum(y_i - \bar{y})^2$
SSR(모형에 의해 설명되는 변동) : $\sum(\hat{y}_i - \bar{y})^2$
SSE(모형에 의해 설명이 되지 않는 변동) : $\sum(y_i - \hat{y}_i)^2$

R^2 = SSR/SST

D

R^2 값	해석
1.0	완벽하게 설명함 (모든 점이 회귀선 위에 있음)
0.7	Y의 변동 중 70%를 X가 설명함
0	아무것도 설명하지 못함

참고) Adj. R-squared (수정된 결정계수)

- 독립변수 개수가 많아질수록 R^2은 자동으로 커지므로,**자유도 보정된 R^2**를 함께 봄.

$$\text{Adj. } R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

n = 표본 크기, k = 독립변수 수

- 변수 추가 시 모델의 성능이 실제로 향상되었는지 판단할 때 유용함.

회귀모델의 종류

- 모형에 포함된 독립변수의 개수에 따라
  - 단순(simple) 회귀: 독립변수가 1개
  - 다중(multiple) 회귀: 독립변수가 2개 이상
- 회귀계수의 형태에 따라
  - **선형(linear) 회귀**
  - 비선형(non-linear) 회귀

회귀모델 결과 해석하기

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        calories_burned   R-squared:                       0.927
Model:                            OLS   Adj. R-squared:                  0.927
Method:                 Least Squares   F-statistic:                 2.554e+04
Date:                Mon, 13 Oct 2025   Prob (F-statistic):               0.00
Time:                        08:16:54   Log-Likelihood:                -53273.
No. Observations:               10000   AIC:                         1.066e+05
Df Residuals:                    9994   BIC:                         1.066e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              2.4310      3.912      0.621      0.534      -5.237      10.099
weight_kg          0.1084      0.033      3.251      0.001       0.043       0.174
steps_per_day      0.0299      0.000    180.938      0.000       0.030       0.030
workout_minutes    5.0208      0.017    296.596      0.000       4.988       5.054
sleep_hours       -1.6849      0.334     -5.039      0.000      -2.340      -1.029
active_minutes     1.9701      0.025     79.177      0.000       1.921       2.019
==============================================================================
Omnibus:                        3.190   Durbin-Watson:                   2.008
Prob(Omnibus):                  0.203   Jarque-Bera (JB):                3.221
Skew:                          -0.041   Prob(JB):                        0.200
Kurtosis:                       2.966   Cond. No.                     6.71e+04
==============================================================================
```

용어	의미	해석 포인트
종속변수 (Dep. Variable)	예측 대상	y
절편 (const)	X=0일 때 Y	
회귀계수 (coef)	X의 영향 크기	1단위 증가 시 Y 변화량
표준오차 (std err)	불확실성	작을수록 신뢰도 높음
**t, P>	t	각 회귀계수의 통계량과 p-value (회귀 계수가 0인지 검정)
[0.025, 0.975]	95% 신뢰구간	계수가 이 구간 내에 있을 확률이 95%
R^2, Adj.R^2	설명력	높을수록 좋음
F-statistic	모델 전체 유의성	전체 회귀모형 검정
Omnibus / Jarque-Bera (JB)	잔차의 정규성 검정	p>0.05면 정규성 만족
Skew / Kurtosis	잔차의 왜도, 첨도	0 근처면 이상치 적음
Durbin-Watson	잔차의 독립성	2에 가까우면 독립성 양호
Cond. No.	다중공선성 진단 지표	30 이상이면 multicollinearity 의심

🎯 다중공선성 (Multicollinearity)

> 다중공선성이란, **여러 독립변수들 간에 강한 선형관계(상관관계)** 가 존재하는 현상을 의미

쉽게 얘기해서 같이 움직이는 변수들을 독립변수에 두었을 때 생기는 현상

- 예: `매출`과 `판매량`, `광고비`와 `노출수`

다중공선성이 높으면 다음과 같은 문제가 발생

⇒ 독립변수들이 서로 비슷한 정보를 가지고 있어 회귀모형이 어떤 변수가 종속변수에 실제로 영향을 주는지 **구분하기 어려워짐.**

문제점	설명
회귀계수 추정의 불안정성	특정 변수를 약간만 변경해도 회귀계수가 크게 변함
변수의 통계적 유의성 저하	t-값이 작아지고, p-value가 높아짐
해석 어려움	각 변수의 영향력을 독립적으로 해석하기 어려움
예측력 왜곡	모델의 일반화 성능이 저하될 수 있음

다중공선성 진단 방법

방법	설명
상관계수 확인	변수 간의 상관계수가 0.8 이상이면 의심
분산팽창지수 (VIF, Variance Inflation Factor)	다중공선성을 수치로 측정하는 대표 지표

VIF 공식

D

- Ri^2: 나머지 변수들로 해당 변수 XiX_iXi를 회귀했을 때의 결정계수
- 일반적으로
  - VIF > 10 → 다중공선성 심각
  - 5 < VIF ≤ 10 → 주의 필요

해결 방법

방법	설명
변수 제거	비슷한 역할을 하는 변수 중 하나 제거
변수 결합	유사한 변수들을 평균·합 등으로 묶기
정규화 회귀 사용	Ridge, Lasso 회귀처럼 규제항을 추가하여 해결
주성분 분석(PCA)	상관된 변수를 축소하여 새로운 독립변수 생성

실습 :)