

선형 회귀 분석

1. 선형 회귀 모델 소개

1. 선형 회귀 모델 소개

1. 선형 회귀 개요

선형 회귀(Linear Regression)는 종속 변수(반응 변수)와 하나 이상의 독립 변수(설명 변수) 간의 선형 관계를 모델링하는 통계적 기법이다. 선형 회귀는 주어진 데이터로부터 직선을 그려서 미래의 값을 예측하는 데 사용된다.

2. 선형 회귀 개념

선형 회귀는 독립 변수 XXX 와 종속 변수 YYY 사이의 관계를 직선 형태로 모델링한다. 단항 선형 회귀(Simple Linear Regression)와 다항 선형 회귀(Multiple Linear Regression)로 나눌 수 있다.

1. 선형 회귀 모델 소개

- **단항 선형 회귀:**

하나의 독립 변수(**x**)와 종속 변수(**y**) 간의 관계를 모델링한다.

$$Y = W * X + b$$

여기서, w 는 기울기(weight), b 는 절편(bias) 이다.

- **다항 선형 회귀:**

여러 독립 변수와 종속 변수 간의 관계를 모델링한다.

$$Y = W_1 X_1 + W_2 X_2 + \dots + W_n X_n + b$$

여기서, b 는 절편, w_1, w_2, \dots, w_n 은 각 독립 변수의 회귀 계수이다.

1. 선형 회귀 모델 소개

3. 최소제곱법

선형 회귀에서 회귀 계수를 추정하기 위해 가장 일반적으로 사용하는 방법은 최소제곱법(Least Squares Method)이다.

최소제곱법은 데이터 포인트와 예측된 직선 사이의 거리(잔차, residual)의 제곱합을 최소화하는 회귀 계수를 찾는 방법이다.

잔차의 제곱합(Residual Sum of Squares, RSS)은 다음과 같이 정의된다:

$$RSS = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

여기서 Y_i 는 실제 값, \hat{Y}_i 는 예측 값이다.

1. 선형 회귀 모델 소개

4. 모델 적합도 평가

선형 회귀 모델의 적합도를 평가하기 위해 다양한 지표가 사용된다

- **결정 계수(R-squared, R^2):** 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표이다. 0과 1 사이의 값을 가지며, 1에 가까울수록 모델이 데이터를 잘 설명한다.

$$R^2 = 1 - \frac{SSR}{SST}$$

여기서, SSR 은 설명된 제곱합(Explained Sum of Squares), SST 는 총 제곱합(Total Sum of Squares)이다.

- **평균 제곱 오차(Mean Squared Error, MSE):** 예측 값과 실제 값 간의 평균적인 차이를 나타내는 지표이다.

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

- **평균 절대 오차(Mean Absolute Error, MAE):** 예측 값과 실제 값 간의 절대적인 차이를 나타내는 지표이다.

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - \hat{Y}_i|$$

5. 선형 회귀 모델의 가정

선형 회귀 모델이 유효하려면 몇 가지 가정이 충족되어야 한다:

- **선형성(Linearity)**: 독립 변수와 종속 변수 간의 관계가 선형이어야 한다.
- **독립성(Independence)**: 관측치들이 서로 독립적이어야 한다.
- **등분산성(Homoscedasticity)**: 모든 수준에서 종속 변수의 분산이 일정해야 한다.
- **정규성(Normality)**: 오차(term)가 정규 분포를 따라야 한다.

6. 다중공선성

다중 선형 회귀에서는 독립 변수들 간의 다중공선성(multicollinearity) 문제가 발생할 수 있다. 다중공선성이란 독립 변수들 간의 상관관계가 높아 회귀 계수의 추정이 불안정해지는 현상을 말한다.

이를 해결하기 위해 다음과 같은 방법들이 사용된다:

- **변수 선택법**: 중요한 변수만을 선택하여 모델을 구축한다.
- **주성분 분석(PCA)**: 변수들을 선형 결합하여 새로운 변수를 생성하고, 이 변수를 사용하여 회귀 분석을 수행한다.
- **정규화 기법**: L1 정규화(Lasso) 또는 L2 정규화(Ridge)를 통해 회귀 계수를 제약하여 다중공선성을 완화한다.

7. 과적합과 과소적합

선형 회귀 모델은 과적합(overfitting)과 과소적합(underfitting) 문제에 주의해야 한다.

- **과적합**: 모델이 학습 데이터에 너무 잘 맞춰져 있어 새로운 데이터에 대한 예측 성능이 떨어지는 현상이다. 복잡한 모델이 과적합을 초래할 수 있다.
- **과소적합**: 모델이 학습 데이터의 패턴을 충분히 학습하지 못해 예측 성능이 떨어지는 현상이다. 지나치게 단순한 모델이 과소적합을 초래할 수 있다.

8. 활용 분야

선형 회귀는 다양한 분야에서 예측 및 데이터 분석에 널리 사용된다:

- **경제학**: 경제 성장 예측, 소비자 지출 예측 등
- **경영학**: 매출 예측, 시장 분석 등
- **사회과학**: 사회 현상 분석, 교육 성과 예측 등
- **의료**: 환자의 치료 반응 예측, 약물 효과 분석 등
- **기타**: 주택 가격 예측, 주식 가격 예측 등

9. 선형 회귀 모델 구축 과정

- 1. 데이터 수집 및 전처리:** 데이터를 수집하고 결측값 처리, 이상치 제거 등의 전처리를 수행한다.
- 2. 데이터 분할:** 데이터를 학습용(training set)과 테스트용(test set)으로 분할한다.
- 3. 모델 학습:** 학습 데이터를 사용하여 선형 회귀 모델을 학습시킨다.
- 4. 예측:** 학습된 모델을 사용하여 테스트 데이터에 대한 예측을 수행한다.
- 5. 모델 평가:** 예측 결과와 실제 값을 비교하여 모델의 성능을 평가한다.

2. 선형 회귀 모델 구현 실습

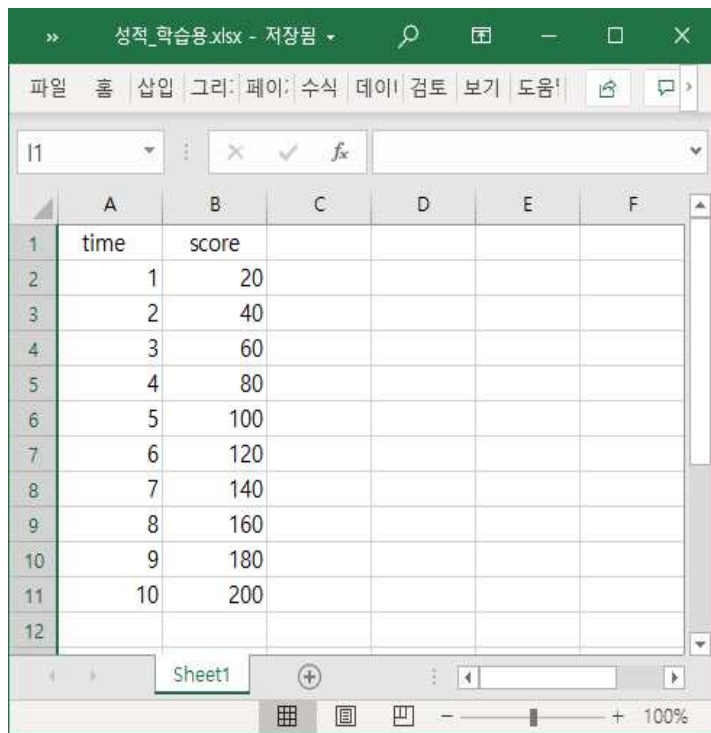
[실습 -1]. 기초 선형 회귀 모델 구현 실습

머신러닝 모델 구현 순서

- 데이터 준비
- 데이터 로딩
- 피처와 학습 모델 선택
- 모델 학습 하기
- 새로운 데이터로 예측하기
- 평가 및 예측결과 보기
- 실제 오차가 있는 데이터로 학습하기

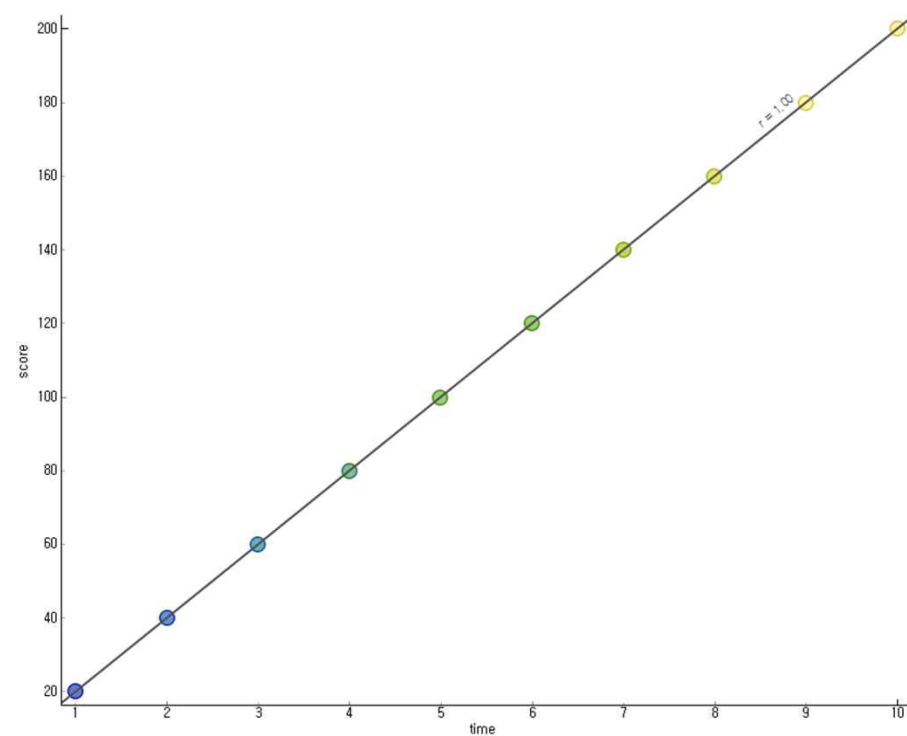
2. 선형 회귀 모델 구현 실습

엑셀을 이용하여 아래와 같이 데이터를 입력하고 파일이름을
'성적_학습용.xlsx'으로 저장해 둔다



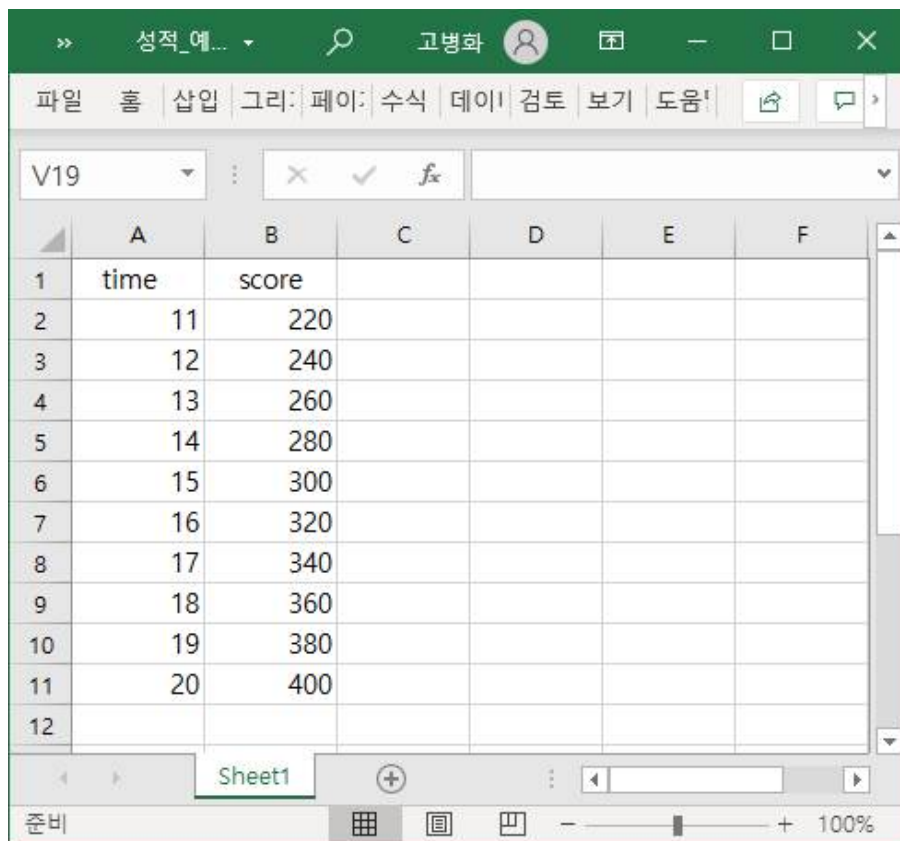
	A	B	C	D	E	F
1	time	score				
2	1	20				
3	2	40				
4	3	60				
5	4	80				
6	5	100				
7	6	120				
8	7	140				
9	8	160				
10	9	180				
11	10	200				
12						

time	score
1	20
2	40
3	60
4	80
5	100
6	120
7	140
8	160
9	180
10	200



2. 선형 회귀 모델 구현 실습

엑셀을 이용하여 아래와 같이 데이터를 입력하고 파일이름을
'성적_예측용.xlsx'으로 저장해 둔다



	A	B	C	D	E	F
1	time	score				
2	11	220				
3	12	240				
4	13	260				
5	14	280				
6	15	300				
7	16	320				
8	17	340				
9	18	360				
10	19	380				
11	20	400				
12						

time	score
11	220
12	240
13	260
14	280
15	300
16	320
17	340
18	360
19	380
20	400

2. 선형 회귀 모델 구현 실습



성적_학습용.xlsx
스프레드시트

📎 데이터 분석해줘(한글로).



업로드하신 파일에는 두 개의 열, ``time``과 ``score``가 포함되어 있습니다. 다음은 데이터를 요약한 내용입니다:

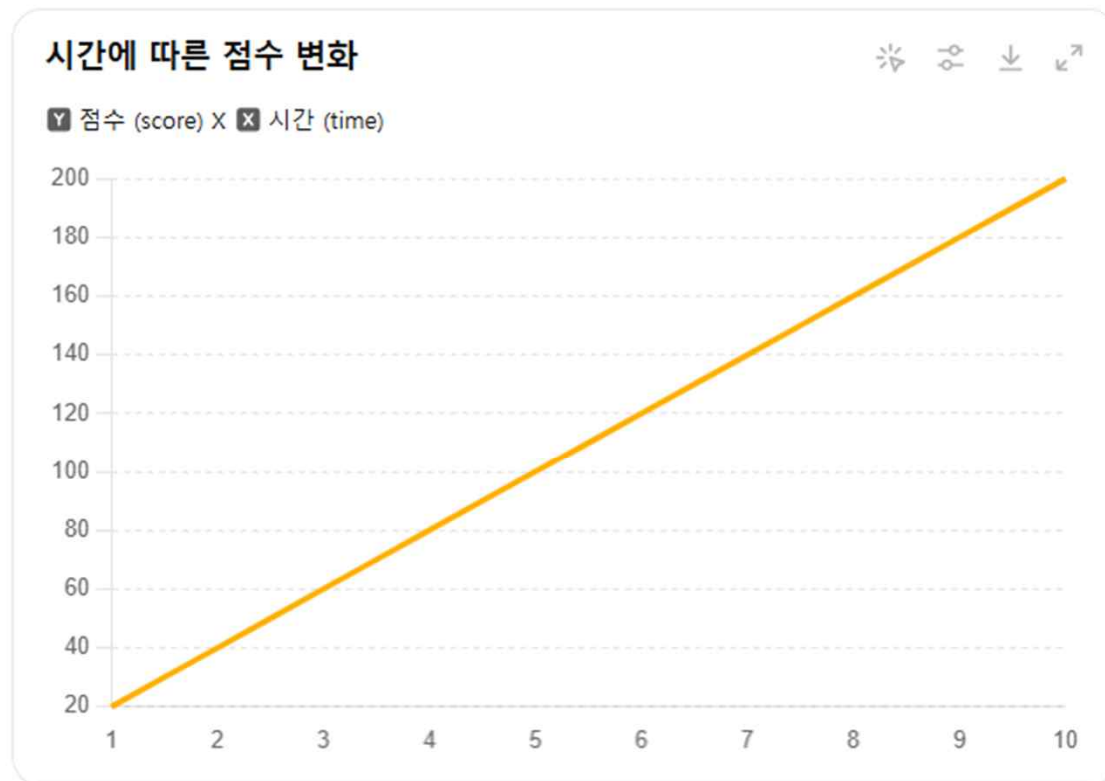
- ``time``: 시간 또는 기간을 나타내는 값입니다.
- ``score``: 해당 시간 또는 기간에 따른 점수입니다.

이 데이터를 바탕으로 몇 가지 분석을 진행해보겠습니다. 어떤 분석을 원하시나요? 예를 들어, 시간에 따른 점수 변화, 평균 점수, 점수의 분포 등을 분석할 수 있습니다. [↗]



2. 선형 회귀 모델 구현 실습

그래프로 출력해줘.



여기 시간에 따른 점수 변화를 시각화한 그래프입니다. 점수가 시간에 따라 증가하는 경향을 보이고 있습니다. 추가로 분석하거나 다른 그래프가 필요하시면 알려주세요. [↗]

2. 선형 회귀 모델 구현 실습

선형 회귀 모델을 만들어주고 결과를 알려줘.

선형 회귀 모델의 결과는 다음과 같습니다:

- 회귀계수 (기울기, **slope**): 20.0
- 절편 (**intercept**): $-4.26e-14$ (거의 0에 가까운 값)
- **R² 값**: 1.0 (이 모델은 데이터를 완벽하게 설명합니다)

이 회귀 모델은 시간이 1 단위 증가할 때마다 점수가 20만큼 증가한다는 것을 나타냅니다. 이 모델을 사용하여 점수를 예측할 수 있습니다. 추가로 보고 싶거나 다른 분석이 필요하시면 알려주세요.

[>]

🔊 📄 ↺ 🔒 ☆

2. 선형 회귀 모델 구현 실습

성적_예측용.xlsx
스프레드시트

이 데이터 셋을 위 모델에 넣어 예측값과 실제값을 표로 출력해주고 평가 지표를 알려줘



이 데이터 셋을 위 모델에 넣어 예측 값과 실제 값을 표로 출력해주고 평가 지표를 알려줘



Predicted Vs Actual Scores			
	time	score	predicted_score
1	11	220	220.000000000000006
2	12	240	240.000000000000006
3	13	260	260.000000000000006
4	14	280	280.000000000000006
5	15	300	300.000000000000006
6	16	320	320.000000000000006

평가 지표는 다음과 같습니다:

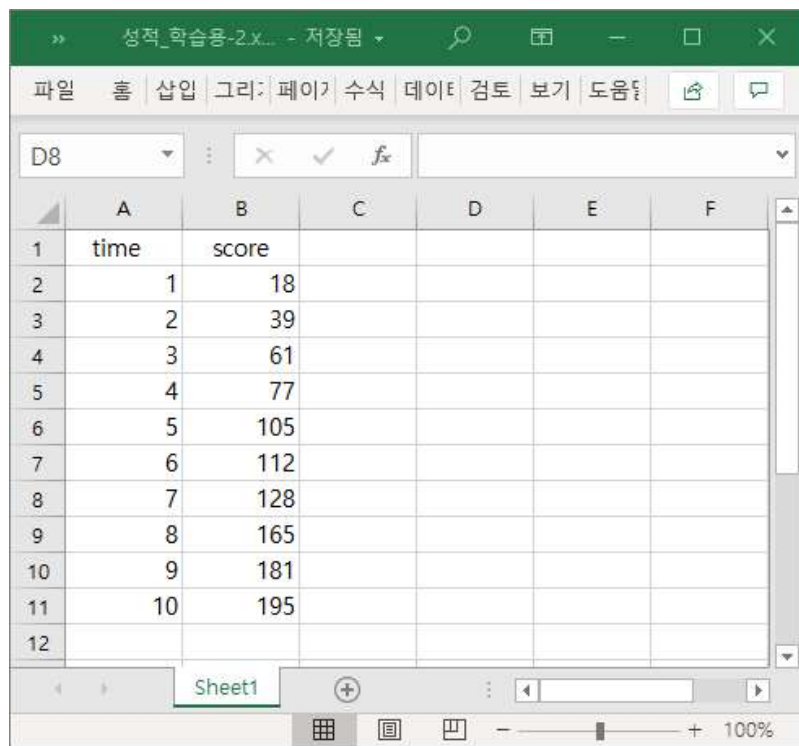
- 평균 제곱 오차 (MSE): 3.231×10^{-27}
- 평균 절대 오차 (MAE): 5.684×10^{-14}
- 평균 제곱근 오차 (RMSE): 5.684×10^{-14}

이 지표들은 모델의 예측이 실제값과 거의 완벽하게 일치함을 나타냅니다.

예측값과 실제값을 비교한 표를 제공합니다.

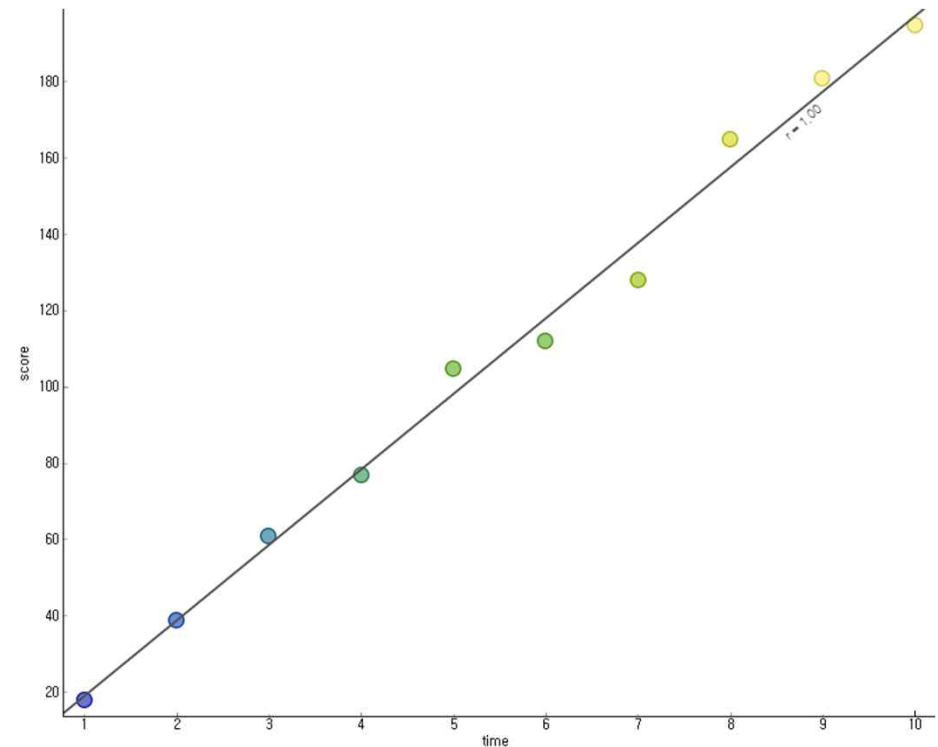
2. 선형 회귀 모델 구현 실습

엑셀을 이용하여 아래와 같이 데이터를 입력하고 파일이름을
'성적_학습용-2.xlsx'으로 저장해 둔다



	A	B
1	time	score
2	1	18
3	2	39
4	3	61
5	4	77
6	5	105
7	6	112
8	7	128
9	8	165
10	9	181
11	10	195

time	score
1	18
2	39
3	61
4	77
5	105
6	112
7	128
8	165
9	181
10	195



2. 선형 회귀 모델 구현 실습



📎 데이터 분석해줘(한글로)



업로드하신 엑셀 파일에는 두 개의 열이 포함된 데이터가 있습니다:

1. ``time`` (시간): 숫자로 표현된 시간 단위
2. ``score`` (점수): 해당 시간에 따른 점수

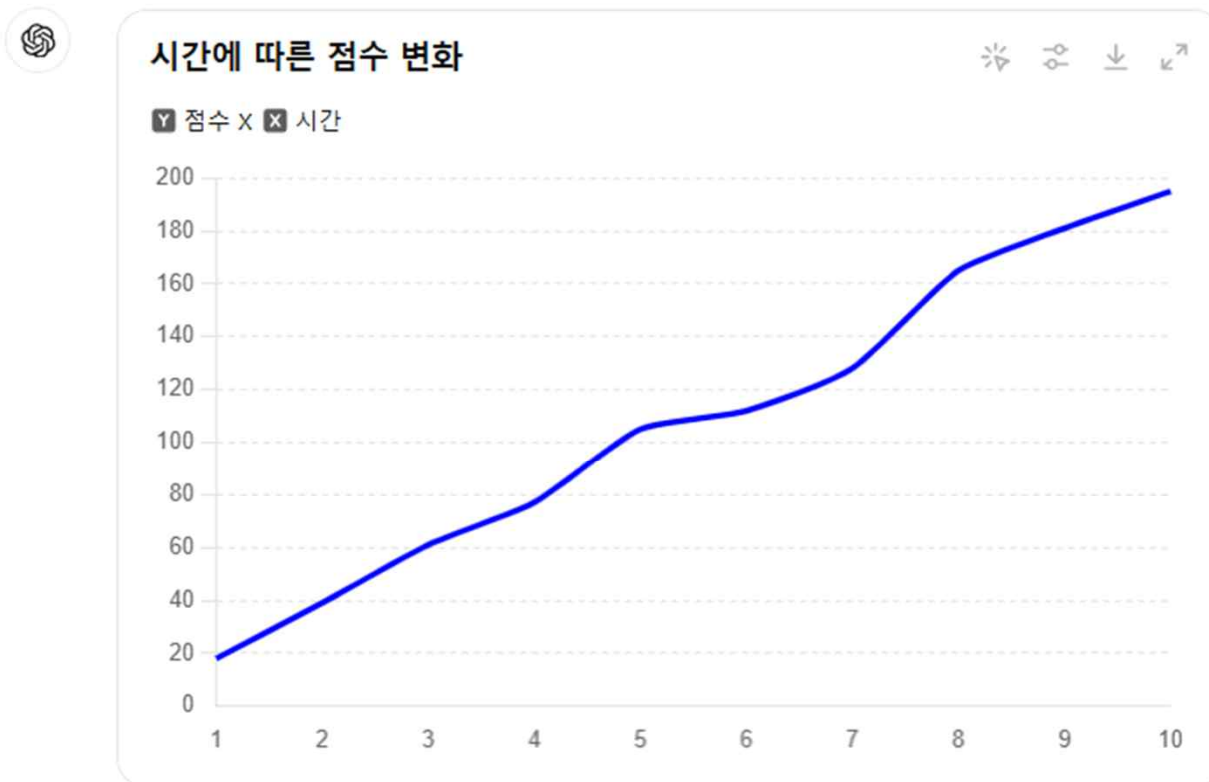
데이터의 예시는 다음과 같습니다:

시간	점수
1	18
2	39
3	61
4	77
5	105

이 데이터는 시간이 지남에 따라 점수가 증가하는 경향을 보여주고 있습니다. 이 데이터를 바탕으로 시간에 따른 점수의 변화를 분석할 수 있습니다. 추가적인 분석이 필요하시면 말씀해 주세요. [↗]

2. 선형 회귀 모델 구현 실습

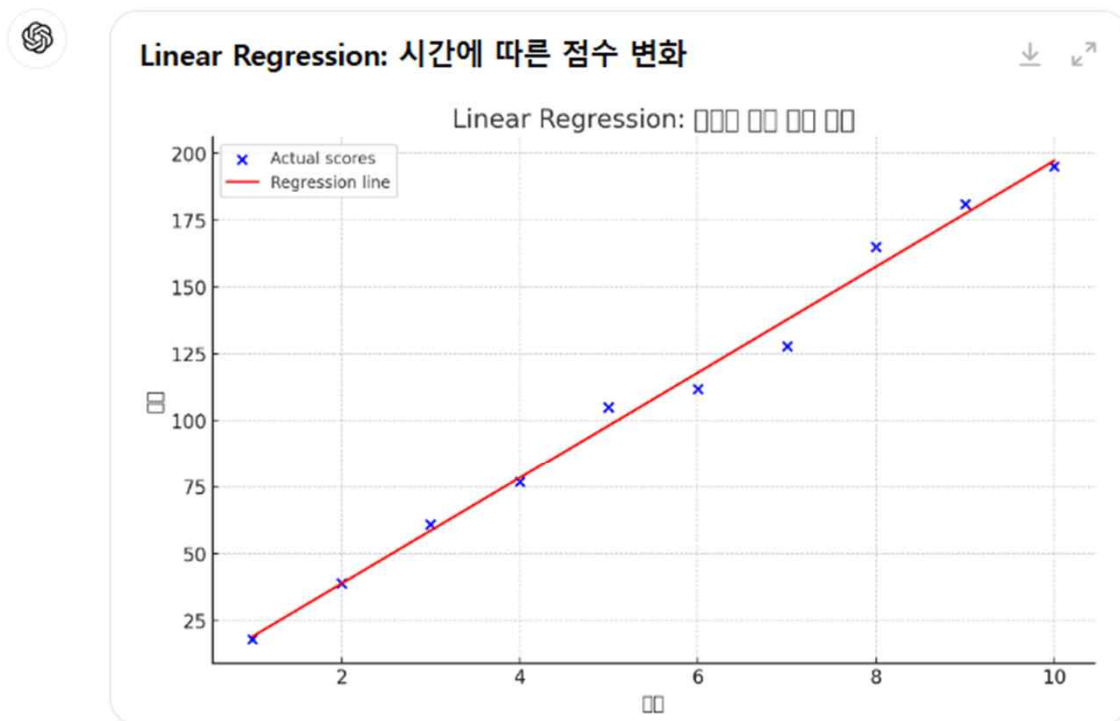
그래프로 출력해줘.



위 그래프는 시간에 따른 점수 변화를 나타냅니다. 시간이 지남에 따라 점수가 증가하는 추세를 볼 수 있습니다. 추가 분석이나 다른 그래프가 필요하시면 알려주세요. [↗]

2. 선형 회귀 모델 구현 실습

선형 회귀모델을 만들어주고 결과를 알려줘.




이 선형 회귀 모델은
점수 = $19.80 * \text{시간} - 0.80$ 의
방정식을 가지고 있습니다.
그래프에서 빨간 선은 회귀선을
나타내며, 파란 점들은 실제
데이터를 나타냅니다

선형 회귀모델의 결과는 다음과 같습니다:

- 절편 (intercept): -0.80
- 기울기 (slope): 19.80

2. 선형 회귀 모델 구현 실습

예측한 값과 실제 값을 비교해서 볼 수 있도록 테이블로 만들어줘.



	Time	Actual Score	Predicted Score	
1	1	18	18.999999999999998	
2	2	39	38.799999999999998	
3	3	61	58.599999999999999	
4	4	77	78.399999999999999	
5	5	105	98.2	
6	6	112	118.0	

예측 값과 실제 값을 비교한 테이블을 생성했습니다. 이를 통해 시간에 따른 점수의 실제 값과 모델이 예측한 값의 차이를 확인할 수 있습니다. 추가적인 분석이 필요하시면 말씀해 주세요. [~]

🔊 📄 ↺ 🔒 ✨

2. 선형 회귀 모델 구현 실습

평가지표도 알려줘.



선형 회귀 모델의 평가 지표는 다음과 같습니다:

- 평균 제곱 오차 (MSE): 25.96
- 결정 계수 (R^2): 0.992

평균 제곱 오차(MSE)는 예측 값과 실제 값 사이의 차이를 제곱하여 평균을 낸 것으로, 값이 작을수록 예측이 실제 값에 가깝다는 것을 의미합니다.

결정 계수(R^2)는 모델의 설명력을 나타내며, 1에 가까울수록 모델이 데이터를 잘 설명한다는 것을 의미합니다. 이번 모델의 R^2 값은 0.992로, 매우 높은 설명력을 가지고 있음을 알 수 있습니다. [↗]



[실습 -2]. 실전 선형 회귀 모델 구현 실습

2. 선형 회귀 모델 구현 실습



boston_train.csv

스프레드시트



데이터 분석해줘(한글로)



ChatGPT는 실수를 할 수 있습니다. 중요한 정보를 확인하세요.

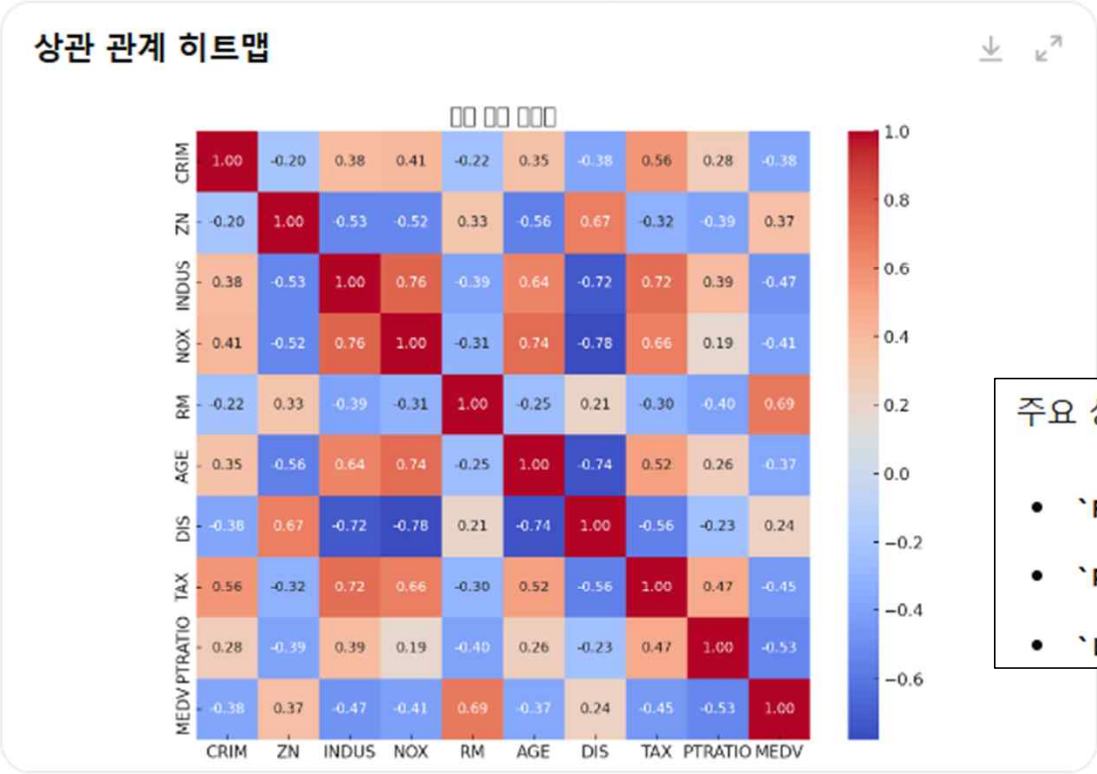


업로드하신 데이터는 보스턴 주택 가격에 대한 데이터입니다. 각 열의 의미는 다음과 같습니다:

- ``CRIM``: 도시 1인당 범죄율
- ``ZN``: 25,000 평방피트를 초과하는 거주 지역의 비율
- ``INDUS``: 비소매 상업 지역의 비율
- ``NOX``: 일산화질소 농도
- ``RM``: 주택당 방의 수
- ``AGE``: 1940년 이전에 건축된 소유 주택의 비율
- ``DIS``: 5개 보스턴 고용 센터까지의 가중 거리
- ``TAX``: 재산세율
- ``PTRATIO``: 학생-교사 비율
- ``MEDV``: 자가 주택의 중앙값 (단위: \$1,000)

2. 선형 회귀 모델 구현 실습

먼저, 상관 관계 행렬을 확인해보겠습니다. [↗]



주요 상관 관계:

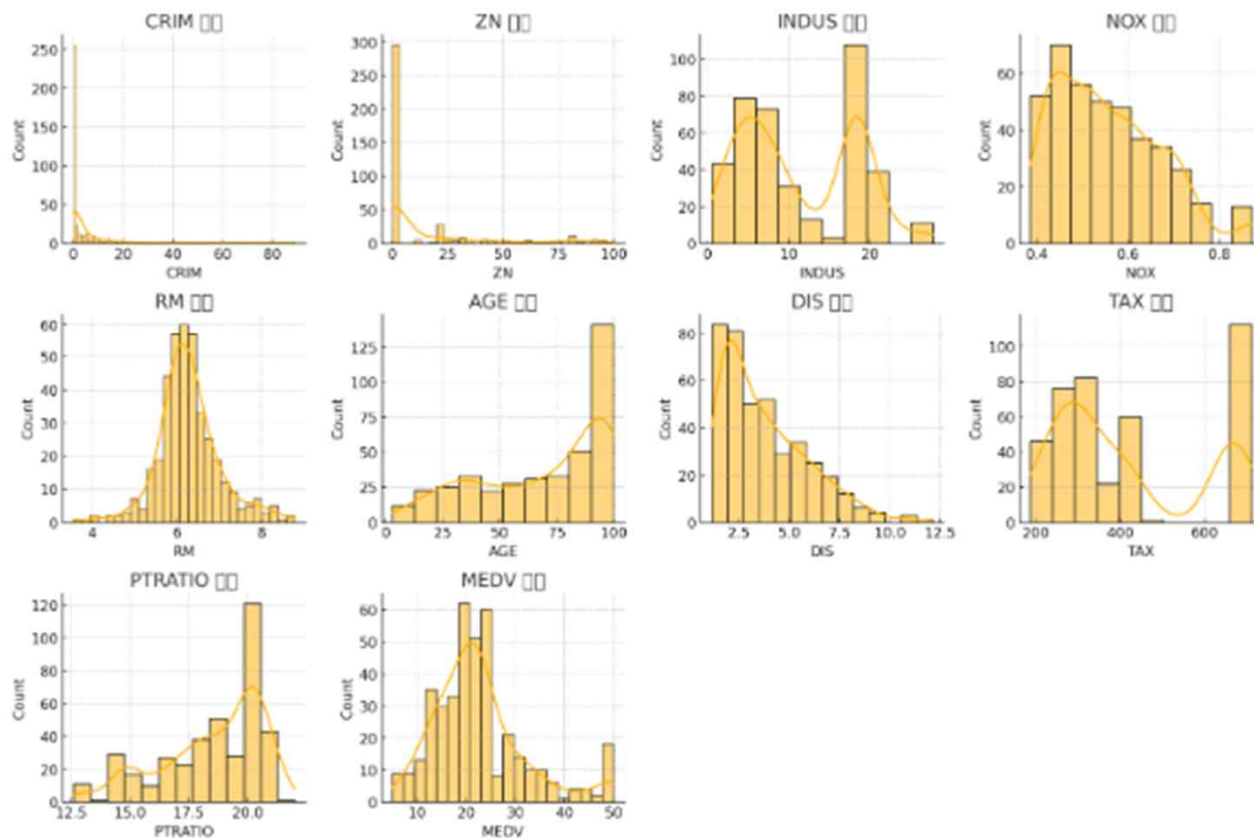
- `RM` (방의 수)과 `MEDV` (주택 가격) 간에는 양의 상관 관계가 존재합니다.
- `PTRATIO` (학생-교사 비율)와 `MEDV` 간에는 음의 상관 관계가 있습니다.
- `LSTAT` (저소득층 비율)과 `MEDV` 간에는 강한 음의 상관 관계가 있습니다.

상관 관계 히트맵을 통해 변수들 간의 상관 관계를 확인할 수 있습니다. 특히, `MEDV`와 다른 변수들 간의 상관 관계가 중요한데, 이 값이 주택 가격의 중앙값을 나타내기 때문입니다.

2. 선형 회귀 모델 구현 실습

이제 주요 변수들의 분포를 히스토그램으로 시각화하겠습니다. [↗]

MEDV 분포



CRIM(범죄율)은 대부분의 값이 낮은 쪽에 집중되어 있고,
RM(방의 수)은 중앙값이 6~7 사이에 위치 합니다.

2. 선형 회귀 모델 구현 실습

‘RM’ 컬럼만 피처로 사용하여 단항 회귀 모델을 만들어줘.



단항 회귀 모델을 통해 ‘RM’ (방의 수)만을 특성으로 사용하여 주택 가격을 예측한 결과는 다음과 같습니다.

모델 계수

- 절편 (Intercept): -33.879
- ‘RM’의 계수 (Coefficient for RM): 9.012

성능 평가

- 학습 데이터
 - MSE (평균 제곱 오차): 50.831
 - R^2 (결정 계수): 0.468
- 테스트 데이터
 - MSE (평균 제곱 오차): 39.242
 - R^2 (결정 계수): 0.475

이를 통해 ‘RM’ (방의 수)이 증가할 때마다 주택 가격이 약 9.012 단위만큼 증가한다는 것을 알 수 있습니다. 모델의 R^2 값은 약 0.47로, 이는 ‘RM’ 변수만으로 주택 가격의 약 47%를 설명할 수 있음을 의미합니다.

2. 선형 회귀 모델 구현 실습



boston_test.csv

스프레드시트

이 데이터 셋으로 위에서 만든 단항 회귀에 넣어 예측하고 평가 지표를 알려줘.



이 데이터 셋으로 위에서 만든 단항 회귀에 넣어 예측하고 평가 지표를 알려줘.



새로운 테스트 데이터셋을 사용하여 단항 회귀 모델의 성능을 평가한 결과는 다음과 같습니다:

- 테스트 데이터 MSE (평균 제곱 오차): 25.685
- 테스트 데이터 R^2 (결정 계수): 0.543

이 결과는 새로운 테스트 데이터셋에서 모델이 비교적 좋은 성능을 보였음을 의미합니다. R^2 값이 약 0.543으로, `RM` (방의 수) 변수만으로 주택 가격의 약 54.3%를 설명할 수 있습니다.

추가적인 분석이나 다른 작업이 필요하시면 말씀해주세요. [↗]

2. 선형 회귀 모델 구현 실습

예측한 값과 실제 값을 비교해서 볼 수 있도록 테이블로 만들어줘.



Actual Vs Predicted MEDV			
	Actual MEDV	Predicted MEDV	
1	24.4	20.767734333682988	
2	36.4	30.80676568271999	
3	21.6	22.588097289253426	
4	28.7	29.914607600534474	
5	33.2	32.951549759085175	
6	18.5	22.47995691565518	

실제 주택 가격(`Actual MEDV`)과 예측된 주택 가격(`Predicted MEDV`)을 비교하는 테이블을 생성하였습니다. 데이터를 직접 확인하실 수 있도록 제공해드렸습니다.

2. 선형 회귀 모델 구현 실습

boston_train.csv
스프레드시트

모든 컬럼을 피쳐로 사용하여 다항 회귀 모델을 만들어줘 |



모든 컬럼을 피쳐로 사용하여 다항 회귀 모델을 만들어줘



모든 컬럼을 피쳐로 사용하여 다항 회귀 모델을 만든 결과는 다음과 같습니다.

모델 계수

- 절편 (Intercept): 29.484
- 각 변수의 계수 (Coefficients):
 - ``CRIM``: -0.156
 - ``ZN``: 0.037
 - ``INDUS``: -0.136
 - ``NOX``: -22.470
 - ``RM``: 6.041
 - ``AGE``: -0.052
 - ``DIS``: -2.001
 - ``TAX``: 0.0007
 - ``PTRATIO``: -1.062

2. 선형 회귀 모델 구현 실습

성능 평가

- 학습 데이터
 - MSE (평균 제곱 오차): 35.026
 - R^2 (결정 계수): 0.633
- 검증 데이터
 - MSE (평균 제곱 오차): 25.798
 - R^2 (결정 계수): 0.655

이 모델은 모든 피쳐를 사용하여 주택 가격을 예측하며, R^2 값이 약 0.655로, 이는 모델이 주택 가격 변동의 약 65.5%를 설명할 수 있음을 의미합니다.

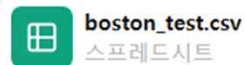
이제 새로운 테스트 데이터셋을 사용하여 예측을 수행하고 성능을 평가하겠습니다. [\[↗\]](#)

새로운 테스트 데이터셋을 사용하여 다항 회귀 모델의 성능을 평가한 결과는 다음과 같습니다:

- 테스트 데이터 MSE (평균 제곱 오차): 17.797
- 테스트 데이터 R^2 (결정 계수): 0.683

이 결과는 모델이 새로운 데이터셋에서 비교적 높은 성능을 보였음을 의미합니다. R^2 값이 약 0.683으로, 이는 모델이 주택 가격 변동의 약 68.3%를 설명할 수 있음을 나타냅니다.

2. 선형 회귀 모델 구현 실습



boston_test.csv

스프레드시트



이 데이터 셋으로 위에서 만든 다항 회귀에 넣어 예측하고 평가 지표를 알려줘.



이 데이터 셋으로 위에서 만든 다항 회귀에 넣어 예측하고 평가 지표를 알려줘.




새로운 테스트 데이터셋을 사용하여 다항 회귀 모델의 성능을 평가한 결과는 이전과 동일하게 나타났습니다:

- 테스트 데이터 MSE (평균 제곱 오차): 17.797
- 테스트 데이터 R^2 (결정 계수): 0.683

이는 모델이 새로운 데이터셋에서도 일관된 성능을 보여주고 있음을 의미합니다.

2. 선형 회귀 모델 구현 실습

예측한 값과 실제 값을 비교해서 볼 수 있도록 테이블로 만들어줘



	Actual MEDV	Predicted MEDV	
1	24.4	22.530043107259203	
2	36.4	34.01505619578056	
3	21.6	23.532083520497824	
4	28.7	31.93486891349981	
5	33.2	34.689240057545696	
6	18.5	26.065883345771496	

실제 주택 가격(`Actual MEDV`)과 예측된 주택 가격(`Predicted MEDV`)을 비교하는 테이블을 생성하였습니다. 데이터를 직접 확인하실 수 있도록 제공해드렸습니다.

1. 선형 회귀 모델 소개

[선형 회귀 모델용 추천 데이터 셋 : <https://archive.ics.uci.edu/>]

- Computer Hardware Data Set : 컴퓨터 하드웨어 상대적 성능 예측
- Wine Quality : Wine 품질 예측
- Solar Flare Data Set : 태양 불꽃 발산 예측
- Forest Fires Data Set : 산불 화재 면적 예측
- Concrete Compressive Strength Data Set : 콘크리트 압축 강도 예측
- Auto MPG Data Set : 자동차 연비 예측(Miles Per Gallon)
- Breast Cancer Wisconsin (Prognostic) Data Set : 위스콘신 대학
유방암 예측
- Real estate valuation data set Data Set : 부동산 가격 예측

감사합니다