회귀 분석:
데이터의 관계 이해

송인규 교수,
지스트 인공지능정책전략대학원 특임교수

통계학이란 무엇인가요?

- 데이터 수집, 분석, 해석에 관한 학문입니다.
- 패턴을 이해하고 정보에 입각한 의사 결정을 내리는 데 도움
- 다양한 분야에서 사용됩니다:
- 비즈니스: 비즈니스: 고객 선호도 이해
- 과학: 가설 테스트
- 스포츠 스포츠: 선수 성과 추적

회귀 분석 소개

- 변수 간의 관계를 이해하는 방법
- 다른 변수를 기반으로 한 변수를 예측하는 데 도움이 됩니다.
- 선형 회귀에 집중합니다:
- 변수 간의 직선적 관계를 가정합니다.
- 예측 및 추세 이해에 유용

변수 유형

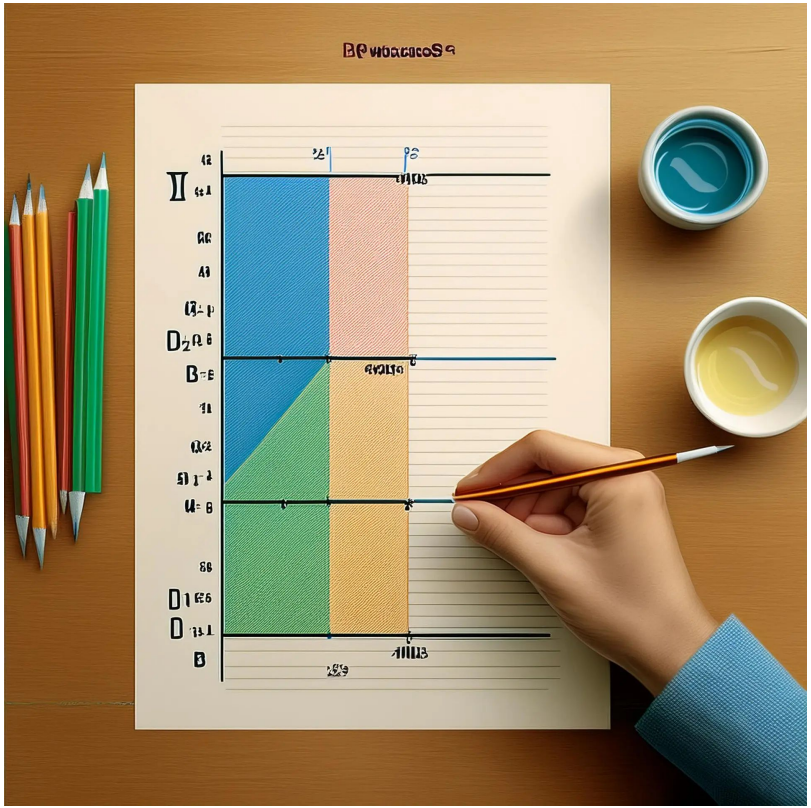
- 종속 변수(응답 변수):
- 예측하고자 하는 대상
- 예시: 예: 시험 점수
- 독립 변수(예측 변수):
- 예측을 위해 사용하는 변수
- 예시: 학습 시간
- 관계: 학습 시간을 기반으로 시험 점수를 예측합니다.

더미 변수

- 범주형 데이터에 사용(숫자가 아닌 데이터)
- 카테고리를 숫자로 변환
- 예시: 학교 유형
- 공립학교 = 0
- 사립 학교 = 1
- 회귀 분석에 범주형 데이터를 포함할 수 있습니다.

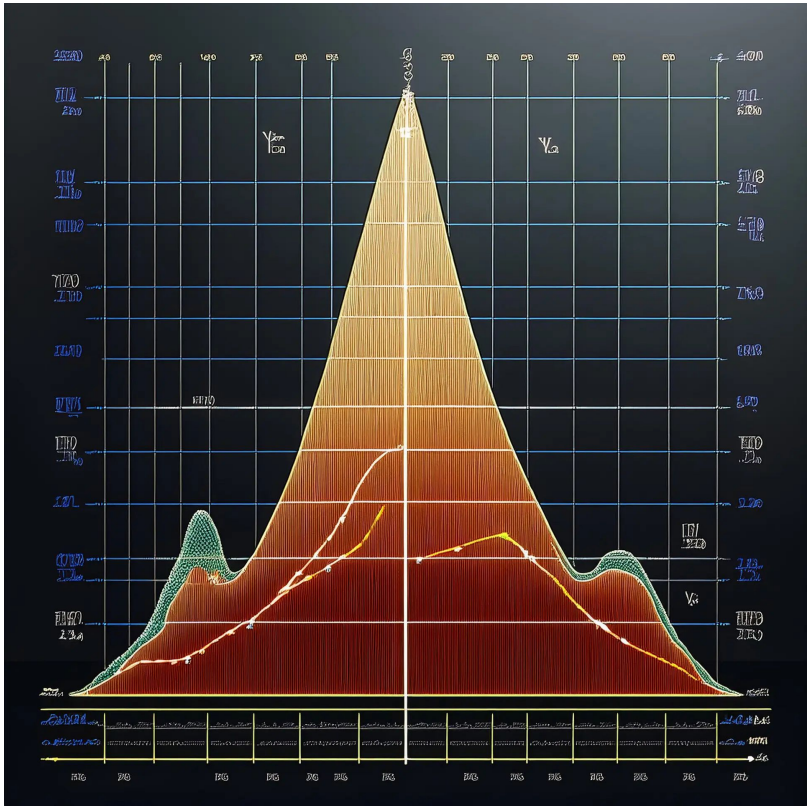
회귀 분석의 기초

- 목표: 데이터 포인트에 가장 잘 맞는 선(회귀선)을 그립니다.
- 이 선은 변수 간의 관계를 보여줍니다.
- 직선의 방정식: $y = mx + b$
- y: 종속 변수
- x: 독립 변수
- m: 기울기
- b: Y-절편



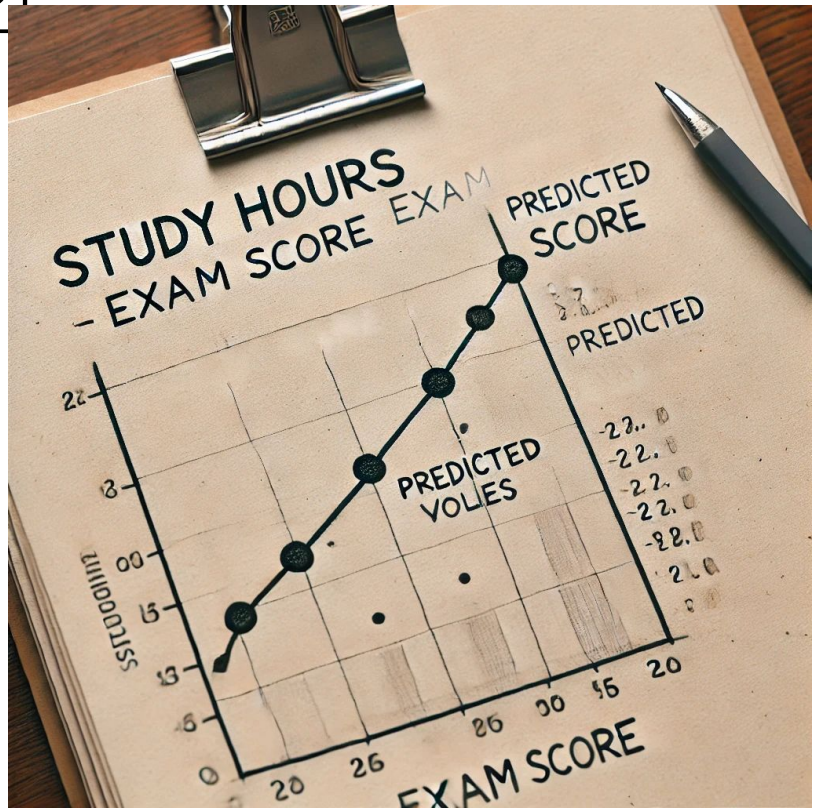
회귀선 해석하기

- 기울기(\(m\)):
- 변화율을 표시합니다.
- 양의 기울기: \(x\)가 증가함에 따라 \(y\)가 증가합니다.
- 음의 기울기: \(x\)가 증가함에 따라 \(y\)가 감소합니다.
- 인터셉트 (\(b\)):
- 선이 Y축을 교차하는 지점
- (x\)가 0일 때 \(y\)의 값입니다.



예시: 시험 점수 및 학습 시간

- 회귀선: 시험 점수 = 10 * 학습 시간 + 50
- 해석:
- 경사 (10): 학습 시간이 추가될 때마다 시험 점수가 10점씩 증가합니다.
- 인터셉트(50): 학습 시간이 0인 경우, 예상 점수는 50점입니다.
- 학습 시간을 기준으로 시험 점수를 예측하는 데 도움이 됩니다.

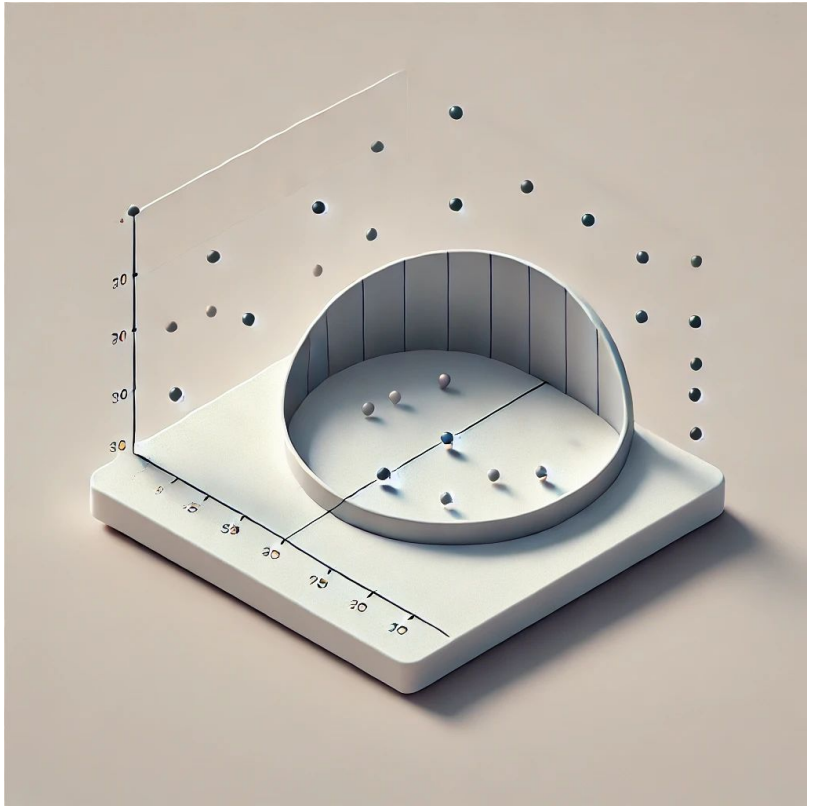


결정 계수(R-제곱)

- 회귀선이 데이터에 얼마나 잘 맞는지 측정합니다.
- 범위는 0에서 1(또는 0%에서 100%)입니다.
- 값이 클수록 더 잘 맞음
- 예시: 예: R-제곱이 0.8이면 종속 변수 분산 중 80%가 독립 변수에 의해 설명된다는 의미입니다.

회귀의 P-값

- 변수 간의 관계가 통계적으로 유의미한지 여부를 표시합니다.
- 낮은 p값(일반적으로 0.05 미만)은 유의미한 결과를 나타냅니다.
- 해석:
- 낮은 p값: 무작위적 우연에 의한 관계가 아닐 가능성이 높음
- p값이 높습니다: 관계는 무작위적인 우연에 의한 것일 수 있음



계수에 대한 신뢰 구간

- 계수(기울기 또는 절편)의 실제 값에 대한 범위를 입력합니다.
- 일반적으로 95% 신뢰 수준에서 계산
- 해석: 실제 계수가 이 범위 내에 있다고 95% 확신합니다.
- 간격이 넓을수록 추정치의 정확도가 떨어집니다.

회귀를 사용하여 예측하기

- 회귀선을 사용하여 새로운 \(x\) 값에 대한 \(y\) 값을 예측합니다.
- 예제 시험 점수 = 10 * 학습 시간 + 50
- 학생이 5시간 동안 공부한 경우:
- 예상 시험 점수 = 10 * 5 + 50 = 100
- 기억하세요: 예측은 추정치이며 보장된 결과가 아닙니다.

예측에 더미 변수 포함하기

- 더미 변수는 카테고리가 결과에 미치는 영향을 보여줄 수 있습니다.
- 더미 변수의 낮은 p값은 유의미한 영향을 시사합니다.
- 예시: 학교 유형(공립/사립)이 낮은 p값인 경우 시험 점수에 큰 영향을 미칩니다.

회귀를 위한 데이터 수집

- 데이터 품질의 중요성:
- 데이터는 관련성이 높고 정확해야 합니다.
- 품질은 예측 정확도에 영향을 미칩니다.
- 오류와 이상값을 피하세요:
- 결과를 왜곡하고 잘못된 결론을 도출할 수 있습니다.
- 분석 전에 데이터를 신중하게 검토하고 정리

더미 변수를 효과적으로 사용하기

- 관련성 있는 카테고리만 포함
- 관련성이 있을 수 있는 요소의 예입니다:
- 학교 유형
- 성별
- 사회경제적 지위
- 종속 변수에 논리적으로 영향을 미칠 수 있는 요인을 선택합니다.

회귀선 그리기

- 일반적으로 정확성을 위해 소프트웨어로 수행
- 단계:
- 데이터 요소의 분산형 차트 만들기
- 최적의 선 계산
- 점을 통해 선 그리기
- 점과 선이 가까울수록 모델 적합도가 높습니다.

회귀 모델 평가하기

- R-제곱 값을 확인합니다:
- 값이 높을수록 적합도가 높음을 나타냅니다.
- 그러나 값이 매우 높으면 과적합을 나타낼 수 있습니다.
- p값을 검토합니다:
- 낮은 p값은 유의미한 변수를 나타냅니다.
- p값이 높으면 모델에서 변수를 제거해야 할 수 있습니다.

선형 회귀의 한계

- 선형 관계만 포착
- 복잡한 비선형 데이터에는 잘 작동하지 않을 수 있습니다.
- 일정한 변화율을 가정함
- 복잡한 데이터에 대해 고려해야 할 다른 모델
- 다항식 회귀
- 다중 회귀
- 비선형 회귀 모델

요약: 회귀 분석의 핵심 사항

- 변수 간의 관계 이해에 도움
- 선형 관계에 \(y = mx + b\) 방정식 사용
- 중요한 메트릭: R-제곱, p-값, 신뢰 구간
- 예측에 유용하지만 제한 사항을 고려해야 합니다.
- 정확한 결과를 얻으려면 데이터 품질과 관련성이 중요합니다.