

October 26, 2025

- 강의명: CS109A: 데이터 과학 입문
- 주차: Lecture 09
- 교수명: Pavlos Protopapas, Kevin Rader, Chris Gumb
- 목적: Lecture 09의 핵심 개념 학습

Part I

개요

강의 핵심 요약

이번 강의는 데이터 과학의 핵심인 **선형 회귀** 모델을 **확률론적 관점**에서 재해석 합니다.

- **왜 확률인가?**: 우리가 가진 데이터는 더 큰 모집단(혹은 데이터 생성 프로세스)에서 나온 하나의 '무작위 실현'일 뿐이므로, 불확실성을 다루기 위해 확률론이 필요합니다.
- **핵심 연결 고리**: 최소제곱법(OLS)을 사용하여 평균제곱오차(MSE)를 최소화하는 것은, 만약 잔차(residuals)가 정규 분포를 따른다고 가정한다면, 통계적 추론 방법인 최대가능도추정(MLE)을 수행하는 것과 수학적으로 동일함을 보입니다.
- **주요 개념**: 확률 변수, PMF/PDF, 정규 분포, 이항 분포, 가능성(Likelihood), 최대가능도추정(MLE)의 기본 원리를 학습합니다.
- **추론 방법 비교**: 모델의 불확실성을 측정하는 두 가지 방법, 즉 공식(t-검정) 기반의 신뢰 구간과 부트스트래핑(Bootstrapping)을 비교합니다.
- **실전 결론**: 모델의 가정(예: 등분산성)이 깨졌을 때, 공식 기반 추론은 잘못된(지나치게 낙관적인) 결론을 내릴 수 있으며, 가정이 적은 부트스트래핑이 더 신뢰할 수 있는 대안이 됩니다.

Part II

핵심 용어 정리

강의에서 다루는 주요 확률 및 통계 용어를 정리합니다.

용어	쉬운 설명(직관)	원어	비고(예시)
확률 변수	어떤 무작위 현상의 결과를 '숫자'로 바꿔주는 변수	Random Variable (RV)	동전 던지기(현상) \rightarrow 앞면=1, 뒷면=0 (RV)
PMF	이산 확률 변수(셀 수 있는)가 특정 값에 가질 확률	Probability Mass Func.	주사위 '3'이 나올 확률 = 1/6
PDF	연속 확률 변수(셀 수 없는)의 '상대 적 가능성'(밀도)	Probability Density Func.	키가 170cm 일 확률은 0이지만, 170-171cm 사이 일 확률은 계산 가능 (곡선 아래 면적)
정규 분포	자연 현상에서 가장 흔히 발견되는 종 모양(bell-shaped)의 연속 분포	Normal (Gaussian) Dist.	사람들의 키, 시험 성적 등
CLT	'많은' 샘플의 '평균'은, 원래 데이터가 어떻든 상관없이, 정규 분포를 따른다는 마법 같은 정리	Central Limit Theorem	모집단이 정규분포가 아니어도 \bar{X} 는 정규분포를 따름
이항 분포	n 번의 독립 시도에서 k 번 성공할 확률(이산 분포)	Binomial Distribution	동전을 10번 던져 앞면이 3번 나올 확률
가능도	'데이터가 관찰된 지금', 어떤 모델(파라미터)이 가장 그럴듯한지에 대한 '믿음의 정도'	Likelihood	$P(\text{Data} \theta)$ 가 아닌 $L(\theta \text{Data})$ 로 관점을 바꾼 것
MLE	가능도를 '최대화'하는 파라미터를 찾는 추정 방법. "이 데이터가 나온 확률이 가장 높은 모델은 무엇인가?"	Max. Likelihood Est.	동전을 10번 던져 앞면이 7번 나왔다면, $p = 0.7$ 을 MLE로 추정
로그 가능도	가능도 함수에 로그(log)를 씌운 것. 미분을 쉽게 하기 위해 사용. (곱셈 \rightarrow 덧셈)	Log-Likelihood	$\log(L(\theta \text{Data}))$
통계적 추론	샘플 데이터를 이용해 모집단의 특성(파라미터)에 대해 추측하는 과정	Statistical Inference	샘플 평균으로 모집단 평균을 추측
표준 오차	추정치의 불확실성(변동성)을 나타내는 값. (추정치의 표준 편차)	Standard Error (SE)	β_1 의 SE: "나의 기울기 추정치가 평균적으로 얼마나 빛나갈까?"
신뢰 구간	모집단 파라미터가 존재할 것이라고 '신뢰'하는 구간(예: 95% 신뢰 구간)	H_0 (귀무가설)과 H_A (대립가설)을 세우고, 데이터로 H_0 을 기각할지 결정하는 절차	Hypothesis Testing
$H_0: \beta_1 = 0$ (효과 없음)			
p-value	귀무가설(H_0)이 '맞다'고 가정할 때, 지금 관찰된 데이터 혹은 더 극단적인 데이터가 나올 확률	p-value	p-value가 낮으면(< 0.05), " H_0 이 맞는데 이런 일이? $\rightarrow H_0$ 을 기각하자"
다중공선성	예측 변수(X)들끼리 강한 상관관계를 보이는 현상	Collinearity	'방의 개수'와 '집 크기(sqft)'가 강하게 비례하는 경우
이분산성	오차(잔차)의 분산이 일정하지 않은 현상	Heteroscedasticity	집 크기가 클수록 가격 예측 오차의 변동 폭도 커짐

Part III

데이터 분석 예시: 주택 가격 예측

강의는 캐임브리지/서머빌 지역의 주택 가격 데이터를 분석하는 예시로 시작합니다.

1 문제 정의 및 데이터 탐색 (EDA)

- 데이터 소스: Redfin.com (온라인 부동산 사이트)
- 데이터 규모: $n = 592$ 개의 주택 판매 기록
- 질문 (목표): 어떤 변수(특성)가 주택 판매 가격과 연관되어 있는가?
- 반응 변수 (Y): price (주택 판매 가격)
- 예측 변수 (X):
 - type: 주택 유형 (콘도, 단독주택, 다가구 등) - (범주형)
 - beds: 침실 수 - (이산형)
 - baths: 욕실 수 - (이산형)
 - sqft: 면적 (제곱피트) - (연속형)
 - lotsize: 대지 크기 - (연속형)
 - year: 건축 연도 - (이산형)
 - dist: 하버드 스퀘어 T-stop(지하철역)까지의 거리 - (연속형)

1.1 데이터 전처리 (Cleaning)

분석 전, 몇 가지 데이터 정리 작업이 수행되었습니다.

1. 결측치 처리 (Missing Data):

- lotsize (대지 크기)와 hoa (주택소유자협회비) 변수에서 많은 결측치 (NA)가 발견되었습니다.
- lotsize는 주로 콘도나 타운하우스에서 결측되었고, hoa는 그 반대였습니다.
- 이는 데이터가 누락된 것이 아니라 '해당 없음'을 의미할 가능성이 높습니다.
- 따라서 이 결측치들을 0으로 대체(Imputation)했습니다. 이는 합리적인 가정입니다.

2. 스케일 조정 (Rescaling):

- price 변수의 단위를 '달러(\$)'에서 '천 달러(\$1000s)'로 변경했습니다.
- 예: \$1,250,000 → 1250.
- 이는 모델의 계수(coefficient)를 해석하기 쉽게 만들습니다.

3. 타입 변환 (Type Conversion):

- zip (우편번호)는 숫자(int)로 저장되어 있었지만, 실제로는 순서나 크기가 없는 범주형 데이터입니다.

- 따라서 문자열(string) 타입으로 변환하여 모델이 이를 연속형 숫자로 오해하지 않도록 했습니다.

2 데이터 시각화 및 주요 발견

데이터 탐색(EDA)을 통해 두 가지 중요한 통계적 문제를 발견했습니다.

2.1 발견 1: 이분산성 (Heteroscedasticity)

- `price`(가격)와 `sqft`(면적)의 산점도(scatter plot)를 확인했습니다.
- 현상: 면적(`sqft`)이 작은 집들은 가격 변동성이 작은 반면(즉, 예측 오차가 적음), 면적이 큰 집들은 가격 변동성이 매우 커집니다(즉, 예측 오차가 큼).
- 정의: 이처럼 예측 변수(X)의 값에 따라 오차(잔차)의 분산이 일정하지 않은 현상을 이분산성(Heteroscedasticity)이라고 합니다.
- 문제점: 이는 표준적인 선형 회귀의 기본 가정('오차의 분산은 일정하다')을 위배합니다. 이 가정이 깨지면, 모델이 추정한 계수($\hat{\beta}$)는 괜찮을지 몰라도, 그 계수의 표준 오차(SE)와 신뢰 구간(CI) 계산이 부정확해집니다.

2.2 발견 2: 다중공선성 (Collinearity)

여러 예측 변수를 모두 포함한 다중 선형 회귀 모델을 피팅했습니다.

다중 회귀 모델 결과 (일부)				
	coef	std err	t	P> t
Intercept	-1949.0670	745.203	-2.615	0.009
sqft	0.6411	0.044	14.720	0.000
beds	-89.9345	23.532	-3.822	0.000
baths	198.4646	31.332	6.334	0.000
...				

- 이상한 점: `beds`(침실 수)의 계수가 음수(-89.9)로 나타났습니다.
- 직관적 해석: "다른 모든 변수(면적, 욕실 수 등)를 고정한 채로, 침실 수만 1개 늘리면 집 값이 약 \$89,934 하락한다."
- 이게 말이 되나?: 상식적으로 침실이 많으면 집값이 올라야 합니다. 이것은 모델의 오류일까요?
- 원인 (다중공선성): 아닙니다. 이는 `beds`와 `sqft` 간의 강한 다중공선성(Collinearity) 때문입니다.
- 올바른 해석 (비유):
 - '다른 모든 변수를 고정한다'는 것은, 특히 '총 면적(`sqft`)을 고정한다'는 의미입니다.

- 즉, 동일한 총 면적의 집에서 침실 수를 1개 늘린다는 것은, 기존의 방들을 쪼개서 더 작고 답답한 침실들을 만든다는 뜻입니다.
- 따라서 ”집이 넓어지지 않는데 방만 얹지로 구겨 넣으면 (cramming another bedroom) 집의 가치가 떨어진다”는 합리적인 해석이 가능합니다.

2.3 모델링 라이브러리: `statsmodels`

- 데이터 과학에서는 `sklearn` 라이브러리를 예측에 주로 사용하지만, 통계적 추론과 해석에는 `statsmodels` 라이브러리가 더 편리할 수 있습니다.
- `statsmodels`는 R 언어처럼 공식(formula)을 사용하여 모델을 정의할 수 있게 해줍니다. (예: "price ~ sqft + type + dist")
- 이는 특히 범주형 변수(type)를 다룰 때 자동으로 더미 변수(dummy variable)를 생성해주는 등 편리함을 제공합니다.

Part IV

복습: 모델 검증 및 규제

본격적인 학률론에 들어가기 전, 이전 강의의 핵심 개념인 교차 검증과 규제를 복습합니다.

3 교차 검증 (Cross-Validation)의 활용

Q: 교차 검증(CV)은 언제 사용하나요?

교차 검증은 특정 모델에 국한된 기술이 아니라, 모델 선택(Model Selection)과 관련된 모든 의사결정에 사용할 수 있는 일반적인 도구입니다.

- A. k-NN 모델에서 최적의 k (이웃 수)를 고를 때
- B. Ridge/Lasso 모델에서 최적의 λ (규제 강도)를 고를 때
- C. 선형 회귀에서 어떤 예측 변수(feature) 조합이 최선인지 고를 때
- D. 서로 다른 모델 계열 (예: k-NN vs. 선형 회귀) 중 어느 것이 더 나은지 비교할 때

정답: A, B, C, D 모두. 이 모든 것은 '모델을 선택'하는 과정이며, CV는 이 선택의 성능을 평가하는 표준 방법입니다.

4 데이터 표준화 (Standardization)

- Q: 예측 변수(X)들을 표준화(Standardize)해야 하는 경우는 언제인가요?
- A: (D) "예측 변수들을 동등하게 처리(treat equally)하고 싶을 때"입니다.
- 이유:
 - k-NN (거리 기반): 표준화를 하지 않으면, '집값'(단위: \$1000s) 같은 큰 스케일의 변수가 '방 개수'(단위: 1) 같은 작은 스케일의 변수보다 거리에 훨씬 큰 영향을 미칩니다. 즉, 스케일이 큰 변수에 편향됩니다.
 - Ridge/Lasso (규제 기반): 규제는 계수(β)의 '크기'에 페널티를 줍니다. 만약 sqft 가 피트 (ft)가 아닌 인치(inch) 단위라면, 스케일이 커져서 계수(β) 값은 0에 매우 가까워질 것입니다. 스케일이 다르면 페널티가 불공평하게 적용되므로, 표준화가 권장됩니다.
- 주의: 표준화는 '항상(Always)' 정답은 아닙니다. 예를 들어, 범주형 변수를 더미(0/1)로 만들었을 때, 이 변수들을 다른 연속형 변수와 동일한 스케일로 맞추는 것이 오히려 모델 해석을 어렵게 하거나 성능을 저하시킬 수도 있습니다.

5 규제 모델 궤적도 (Trajectory Plots) 해석

Lasso와 Ridge 모델에서 규제 강도(λ 또는 α)를 변화시킬 때, 각 변수의 회귀 계수(β)가 어떻게 변하는지 그런 그래프입니다.

Lasso vs. Ridge 케적도

- **Lasso (라쏘):** λ 가 커지면 계수가 정확히 0이 됩니다. 0이 된 변수는 모델에서 '탈락'한 것이므로, 특성 선택(Feature Selection)에 유용합니다.
 - **Ridge (릿지):** λ 가 커져도 계수가 0에 가까워질 뿐, 정확히 0이 되지는 않습니다. 모든 변수를 유지하되 영향력을 줄입니다.
-
- **이상적인 케적도:** 교과서에 나오는 케적도는 매우 매끄러운 곡선을 그리며 0으로 수렴합니다. 이는 모든 예측 변수(X)들이 서로 독립(independent)이라고 가정한, 비현실적인 상황입니다.
 - **현실적인 케적도:** 실제 데이터에서는 다중공선성(Collinearity) 때문에 케적도가 매우 지저분합니다.
 - **다중공선성의 징후:**
 - 특정 계수가 0으로 수렴하다가 갑자기 부호가 바뀌거나 (0을 통과함)
 - 0으로 수축하는 대신 오히려 일시적으로 값이 커졌다가 줄어드는 경우
 - 이는 λ 가 커짐에 따라 한 변수(A)가 페널티를 받아 영향력이 줄어들 때, 그와 상관관계가 높은 다른 변수(B)가 A의 예측력(power)을 '넘겨받아' 계수가 커지는 현상을 나타냅니다.

Part V

확률론의 기초

데이터 과학 모델의 근간이 되는 확률 이론의 기본 개념들을 복습합니다.

6 확률과 확률 변수

- 확률(Probability)이란?
 - 정의: 어떤 사건이 발생할 장기적인 상대 빈도(long-run, relative frequency).
 - 범위: 0 (절대 발생 안 함)에서 1 (항상 발생) 사이의 값을 가집니다.
- 왜 데이터 과학에서 확률이 중요한가?
 - 우리가 가진 데이터(샘플)는 더 큰 데이터 생성 프로세스(Data Generating Process) 또는 모집단에서 나온 하나의 무작위적인 실현(random realization)에 불과합니다.
 - 확률론은 이 '불확실성'을 수학적으로 다루고, 샘플을 넘어선 모집단에 대한 추론을 가능하게 하는 언어입니다.
- 확률 변수(Random Variable, RV)란?
 - 무작위적인 현상(phenomenon)의 결과를 숫자(numeric outcome)로 대응시키는 함수(또는 변수)입니다.
 - 예시: "하버드 학생이 Mac 노트북을 사용하는가?"라는 현상을 조사할 때
 - $X_1 =$ 첫 번째 학생의 응답
 - $X_1 = 1$ (Mac 사용자), $X_1 = 0$ (그 외 사용자) $\rightarrow X_1$ 은 확률 변수입니다.

7 핵심 구분: PMF vs. PDF (이산형 vs. 연속형)

확률 변수는 크게 두 종류로 나뉘며, 이에 따라 확률을 기술하는 방식이 달라집니다.

이산형 확률 변수 (Discrete RV)

- 정의: 변수가 가질 수 있는 값이 '셀 수 있는' 경우 (예: 0, 1, 2, ... 또는 정수 값).
- 확률 함수: 확률 질량 함수 (Probability Mass Function, PMF).
- 특징: $P(X = x)$ 값이 특정 '확률'을 가집니다. 막대 그래프(bar chart)로 시각화됩니다.
- 예시: 동전 던지기 (0, 1), 주사위 굴리기 (1, 2, 3, 4, 5, 6), 침실 수 (1, 2, 3, ...).

연속형 확률 변수 (Continuous RV)

- 정의: 변수가 특정 범위 내의 '모든 실수' 값을 가질 수 있는 경우.
- 확률 함수: 확률 밀도 함수 (Probability Density Function, PDF).
- 특징:
 - $f(x)$ 값(곡선의 높이)은 확률이 아니라 '밀도(density)' 또는 '상대적 가능성'입니다.
 - 특정 값에서의 확률은 0입니다. (예: $P(\kappa = 175.000\ldots\text{cm}) = 0$).
 - 확률은 항상 '구간'으로 계산되며, 이는 PDF 곡선 아래의 면적(Area)과 같습니다.
- 예시: 사람의 키, 주택 가격(price), 지하철 역까지의 거리(dist).

8 주요 확률 분포

8.1 베르누이 분포 (Bernoulli Distribution)

- 정의: '단 한 번'의 시도($n = 1$)에서 '성공(1)' 또는 '실패(0)' 두 가지 결과만 나오는 분포. (가장 단순한 이산 분포)
- 파라미터: p (성공할 확률)
- 예시: 동전 1번 던지기 (앞면=1, 뒷면=0), 학생 1명이 Mac 유저(1)인지 아닌지(0).
- PMF: $P(X = x) = p^x(1 - p)^{1-x}$, ($\text{단}, x \in \{0, 1\}$)

8.2 이항 분포 (Binomial Distribution)

- 정의: '서로 독립'인 베르누이 시도를 n 번 반복했을 때, '성공 횟수(x)'를 나타내는 이산 분포.
- 파라미터: n (총 시도 횟수), p (각 시도의 성공 확률)
- 예시: 동전을 10번($n = 10$) 던졌을 때, 앞면이 3번($x = 3$) 나올 확률 (단, $p = 0.5$).
- PMF: $P(X = x) = \binom{n}{x} p^x(1 - p)^{n-x}$
- $\binom{n}{x}$ (n choose x)의 의미: " n 번의 시도 중 성공이 x 번 발생하는 모든 경우의 수"를 의미합니다. (예: HHH...TTT, HTHTH... 등). 이항 계수(Binomial Coefficient)라고 부릅니다.

8.3 정규 분포 (Normal Distribution)와 중심극한정리 (CLT)

- 정의: 자연계와 사회 현상에서 가장 흔하게 발견되는 종 모양(bell-shaped)의 연속 분포. 가우시안 분포(Gaussian Distribution)라고도 합니다.
- 파라미터: μ (평균, 분포의 중심), σ^2 (분산, 분포의 평균 정도)
- 표준 정규 분포 (Standard Normal): 평균이 0이고 표준편차가 1인($Z \sim N(0, 1)$) 특별한 정규 분포.
- 표준화 (Standardization): 어떤 정규 분포 $X \sim N(\mu, \sigma^2)$ 라도, $Z = \frac{X-\mu}{\sigma}$ 공식을 통해 표준 정규 분포로 변환할 수 있습니다. Z 값은 "평균에서 몇 표준편차만큼 떨어져 있는가?"를 의미 합니다.

핵심 원리: 중심극한정리 (Central Limit Theorem)

Q: 왜 정규 분포가 이렇게 중요한가요?

A: 중심극한정리(CLT) 때문입니다.

CLT란? ”모집단이 어떤 분포(정규분포가 아니어도 됨)를 따르든 상관없이, 거기서 뽑은 샘플의 크기 n 이 충분히 크다면, 그 샘플들의 평균(\bar{X})이 이루는 분포는 정규 분포에 근사한다.”

- **직관적 예시 (신장):** 한 사람의 키(Height)는 수많은 작은 요인들(유전, 영양, 수면, ...)의 '합' 또는 '평균'으로 결정됩니다. CLT에 의해, 이렇게 많은 요인들이 합쳐진 결과물은 정규 분포를 따르게 됩니다.
- **수학적 표현:** $X_i \sim (\mu, \sigma^2)$ 일 때, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- **중요한 함의:**
 - 샘플 평균(\bar{X})의 평균은 모집단 평균(μ)과 같습니다.
 - 샘플 평균(\bar{X})의 분산은 n 이 커질수록 작아집니다 (n 으로 나눔).
 - (직관: 샘플 1개를 뽑는 것보다, 100개를 뽑아 평균내는 것이 훨씬 더 안정적이고 실제 평균에 가깝습니다.)

Part VI

최대가능도추정 (MLE)과 선형 회귀의 연결

이번 강의의 가장 핵심적인 부분으로, OLS 회귀 모델이 어떻게 확률론적 MLE와 연결되는지 설명합니다.

9 추론: 확률의 역방향 문제

확률과 통계적 추론은 동전의 양면과 같습니다.

확률 (Probability) vs. 추론 (Inference)

- **확률 (Deduction):** 모델(파라미터) → 데이터
 - 질문: ”공정한 동전($p = 0.5$)이 주어졌을 때, 10번 던져 앞면이 8번 나올 확률 ($P(\text{Data}|\text{Model})$)은 얼마인가?”
 - 방향: 원인 → 결과
- **통계적 추론 (Inference):** 데이터 → 모델(파라미터)
 - 질문: ”동전을 10번 던져 앞면이 8번 나왔다(Data). 이 동전은 공정한가($p = 0.5$), 아니면 편향되었는가($p = 0.8$)? 어떤 모델이 이 데이터를 가장 잘 설명하는가?”
 - 방향: 결과 → 원인 (우리가 데이터 과학에서 하는 일!)

10 가능성 함수 (Likelihood Function)

- 정의: 가능성 함수 $L(\theta|\text{data})$ 는 PMF 또는 PDF와 수학적으로는 동일한 함수입니다.
- 관점의 차이:
 - PDF/PMF, $f(\text{data}|\theta)$: θ (파라미터)를 고정하고 data를 변수로 봅니다.
 - Likelihood, $L(\theta|\text{data})$: data를 (우리가 관찰한) 고정된 값으로 보고, θ (파라미터)를 변수로 봅니다.
- 의미: 이 함수는 ”관찰된 데이터 하에서, 이 파라미터 θ 가 얼마나 그럴듯한지(likely)”를 측정 합니다.
- 독립 가정: 만약 n 개의 데이터가 모두 독립적으로 (i.i.d.) 샘플링되었다면, 전체의 가능성은 각 데이터 포인트의 가능성의 곱(product)으로 표현됩니다.

$$L(\theta|x_1, \dots, x_n) = \prod_{i=1}^n L(\theta|x_i) = \prod_{i=1}^n f(x_i|\theta)$$

11 로그 가능도 (Log-Likelihood) 와 MLE

- 문제점: 수많은 확률값을 '곱하는' 것은 수학적으로 매우 다루기 어렵습니다. (값이 0에 가깝게 작아지거나, 미분이 복잡해짐)
- 해결책: 로그 가능도 (Log-Likelihood) $l(\theta) = \log(L(\theta))$ 를 사용합니다.
- 이유:
 - log 함수는 단조 증가(monotonic) 함수이므로, $L(\theta)$ 를 최대화하는 θ 값은 $l(\theta)$ 를 최대화하는 θ 값과 동일합니다.
 - 로그의 성질($\log(A \times B) = \log(A) + \log(B)$) 덕분에, 거대한 곱셈(Product)이 간단한 덧셈(Sum)으로 바뀝니다.

$$l(\theta|\text{data}) = \log \left(\prod_{i=1}^n f(x_i|\theta) \right) = \sum_{i=1}^n \log(f(x_i|\theta))$$

- 최대가능도추정 (Maximum Likelihood Estimator, MLE):
 - 정의: 이 로그 가능도 함수 $l(\theta)$ 를 최대화하는 파라미터 $\hat{\theta}$ 를 찾는 방법입니다.
 - 직관: "내가 가진 데이터를 만들어 냈을 가능성인 가장 높은 파라미터(모델)를 찾겠다!"
 - 예시: 데이터 $[3, 5, 10]$ 이 $N(\mu, \sigma^2 = 4)$ 에서 나왔다고 가정할 때, 이 데이터의 로그 가능도 함수는 $\mu = 6$ 일 때 최대가 됩니다. 따라서 MLE $\hat{\mu} = 6$ 이며, 이는 샘플 평균(\bar{x})과 같습니다.
- MLE를 찾는 방법:
 - 분석적 방법 (Analytical): $l(\theta)$ 를 θ 에 대해 미분하고, 그 값을 0으로 만드는 θ 를 찾습니다. (수학적으로 해가 구해지는 경우)
 - 수치적 방법 (Numerical): 해가 복잡할 때 컴퓨터를 사용합니다. 경사 하강법(Gradient Descent)을 이용해 '음의' 로그 가능도 (Negative Log-Likelihood)를 최소화(minimize)합니다. (최대화 \rightarrow 음수 최소화)

12 OLS와 MLE의 통합 (The Big Connection)

이제, 왜 우리가 선형 회귀에서 손실 함수로 평균제곱오차(MSE)를 사용했는지에 대한 확률론적 정당성을 찾게 됩니다.

핵심 결론: OLS는 MLE의 특별한 경우이다

가정: 선형 회귀 모델 $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ 에서, 잔차(error term) ϵ_i 가 서로 독립이며 평균이 0인 정규 분포 $\epsilon_i \sim N(0, \sigma^2)$ 를 따른다고 가정하자.

1. 모델 재정의: 위 가정은 Y_i 자체가 X_i 에 조건부로 정규 분포를 따른다는 의미입니다.

$$Y_i | X_i \sim N(\text{mean} = \beta_0 + \beta_1 X_i, \text{variance} = \sigma^2)$$

2. 가능성 함수 작성: n 개의 모든 데이터에 대한 (로그) 가능성 함수를 작성합니다. 정규 분포 PDF 공식을 사용하고 log를 써워 덧셈으로 만듭니다.

$$l(\beta_0, \beta_1, \sigma^2 | \text{Data}) = \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma} \right)^2} \right)$$

3. 함수 정리: 로그를 풀면 두 부분으로 나뉩니다.

$$l(\dots) = \underbrace{\sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)}_{(A) \beta \text{ 와 무관한 상수항}} - \underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2}_{(B) \beta \text{ 가 포함된 항}}$$

4. MLE 목표: 우리의 목표는 이 $l(\dots)$ 함수를 최대화하는 β_0 와 β_1 를 찾는 것입니다 (MLE).

5. 결론:

- (A) 부분은 β_0, β_1 와 아무 상관이 없으므로 무시할 수 있습니다.
- $l(\dots)$ 를 최대화하려면, (B) 부분을 최대화해야 합니다.
- (B) 앞에는 음수(-) 부호가 붙어 있으므로, (B)를 최대화하는 것은 괄호 안의 $\sum(\dots)^2$ 부분을 최소화하는 것과 같습니다.
- $\sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \leftarrow$ 이것이 바로 오차제곱합 (Sum of Squared Errors, SSE)입니다!

최종 요약: '잔차가 정규 분포를 따른다'고 가정한 상태에서 MLE를 찾는 것은, 수학적으로 OLS(최소제곱법)를 사용하여 SSE(혹은 MSE)를 최소화하는 것과 정확히 일치합니다. 이는 우리가 왜 손실 함수로 MSE를 사용하는지에 대한 강력한 이론적 근거를 제공합니다.

Part VII

통계적 추론: 불확실성 정량화

모델이 $\hat{\beta}_1 = 0.5898$ 이라는 '하나의 값'(점 추정치)을 주었지만, 이 추정치가 얼마나 신뢰할 수 있는지(불확실성)를 알아야 합니다.

13 점 추정(Point Estimate)의 한계

- 우리가 얻은 $\hat{\beta}_1 = 0.5898$ 은 $n = 592$ 개의 '샘플'로 계산한 값입니다.
- 만약 우리가 다른 592개의 샘플을 뽑았다면, $\hat{\beta}_1$ 값은 0.59나 0.57처럼 약간 다른 값이 나왔을 것입니다.
- 즉, 우리의 추정치($\hat{\beta}$) 자체도 '불확실성'을 가집니다.
- 통계적 추론의 목표는 이 불확실성을 정량화하는 것입니다. (예: "진짜 β_1 이 0.6일 가능성은?", "0일 가능성은?")
- 우리는 두 가지 방법(신뢰 구간, 가설 검정)을 사용합니다.

14 신뢰 구간(Confidence Interval, CI)

- 목표:** "진짜 β_1 (모집단의 기울기)이 존재할 것이라고 95% 신뢰하는 구간을 제공하자."
- 방법 1 (부트스트래핑, Bootstrapping):** (이전 강의에서 배움)
 - 원본 데이터(592개)에서 중복을 허용하여 592개를 다시 뽑습니다(Resampling).
 - 모델을 다시 피팅하여 $\hat{\beta}_1^*$ (부트스트랩 추정치)를 기록합니다.
 - 이 과정을 1000번 반복하여 $\hat{\beta}_1^*$ 의 분포를 만듭니다.
 - 이 분포의 하위 2.5%와 상위 97.5% 지점을 찾아 95% 신뢰 구간으로 삼습니다.
- 방법 2 (공식 기반, Formula-based):** (이번 강의)
 - 확률 이론(CLT, t-분포)을 바탕으로 신뢰 구간을 계산하는 공식을 사용합니다.
 - 신뢰 구간 공식: 추정치 \pm (임계값) \times (표준 오차)
 - $$\hat{\beta}_1 \pm t^* \cdot \hat{SE}(\hat{\beta}_1)$$

14.1 표준 오차(Standard Error, SE)

- 정의:** 표준 오차($SE(\hat{\beta}_1)$)는 추정치($\hat{\beta}_1$)의 표준 편차입니다.
- 직관:** "우리의 추정치가 평균적으로 얼마나 부정확한가?" (불확실성의 크기)
- 선형 회귀의 가정(정규성, 등분산성 등)이 모두 맞는다면, $\hat{SE}(\hat{\beta}_1)$ 를 계산하는 수학 공식이 존재합니다.

15 가설 검정 (Hypothesis Testing)

- 목표: ”특정 가설(예: ’sqft는 가격에 영향이 없다’)이 맞는지 틀리는지 데이터로 검증하자.”
- 단계 (t-검정 예시):
 1. 가설 설정:
 - H_0 (귀무가설): $\beta_1 = 0$ (sqft는 가격과 관계가 없다. 기울기는 0이다.)
 - H_A (대립가설): $\beta_1 \neq 0$ (sqft는 가격과 관계가 있다. 기울기는 0이 아니다.)
 2. 검정 통계량 (Test Statistic) 계산:
 - $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
 - 직관: ”우리가 관찰한 기울기 ($\hat{\beta}_1$) 가, 0으로부터 몇 표준 오차(SE) 만큼 떨어져 있는가?”
 - t 값이 크다는 것은 $\hat{\beta}_1$ 이 0과 매우 멀리 떨어져 있다는 뜻이며, 이는 H_0 에 불리한 증거입니다.
 3. p-value 계산:
 - ” H_0 이 정말 맞다($\beta_1 = 0$)고 가정할 때, 지금 우리가 관찰한 t 값만큼 (혹은 더) 극단적인 t 값이 우연히 나올 확률은 얼마인가?”
 4. 결정:
 - p-value가 매우 낮으면 (관례적으로 < 0.05), ”우연이라고 보기엔 너무 드문 일이다. H_0 이 틀린 것 같다.” → H_0 을 기각(Reject) 합니다.
 - ”sqft는 가격에 통계적으로 유의 미한(statistically significant) 영향을 미친다”고 결론 내립니다.

16 부트스트래핑 vs. 공식: 최종 비교

주택 가격 예시(price ~ sqft)로 돌아가, 두 가지 방법으로 계산된 β_1 (sqft의 계수)의 95% 신뢰 구간을 비교합니다.

가정 위반의 결과: 부트스트래핑의 승리

- 방법 1 (부트스트래핑 결과):
 - CI: [0.487, 0.705]
 - 구간의 너비: $0.705 - 0.487 = 0.218$
- 방법 2 (공식 기반 `statsmodels` 결과):
 - CI: [0.544, 0.636]
 - 구간의 너비: $0.636 - 0.544 = 0.092$

결과 분석:

1. 공식 기반의 신뢰 구간이 부트스트래핑 신뢰 구간보다 훨씬 좁습니다(narrower).
2. 이는 공식 기반 방법이 ”우리의 추정치가 매우 정확하다”고 지나치게 낙관적 인(overly optimistic) 결론을 내렸음을 의미합니다.
3. 왜 이런 일이 발생했는가?
 - 공식 기반 방법은 선형 회귀의 모든 가정이 완벽하게 충족될 때만 정확합니다.
 - 하지만 우리는 EDA 과정에서 이분산성(Heteroscedasticity)을 발견했습니다. 이는 ’오차의 분산이 일정하다’는 핵심 가정을 위반한 것입니다.
 - 가정이 깨졌기 때문에, 공식을 통해 계산된 표준 오차(\hat{SE}) 값이 잘못(너무 작게) 계산된 것입니다.
4. 최종 결론:
 - 부트스트래핑은 데이터의 실제 분포(이분산성 포함)를 그대로 사용하여 추정치의 불확실성을 계산합니다.
 - 따라서 회귀 모델의 가정이 위반되었을 때, 부트스트래핑이 공식 기반 방법보다 더 정직하고 신뢰할 수 있는(reliable) 불확실성 추정치를 제공합니다.

Part VIII

빠르게 훑어보기 (1페이지 요약)

1. 핵심 연결: OLS

최소제곱법(OLS)을 사용하여 MSE(오차제곱합)를 최소화하는 것은, 만약 ”잔차가 정규 분포를 따른다”고 가정한다면, 확률론적인 최대가능도추정(MLE)을 수행하는 것과 수학적으로 완벽하게 동일합니다.

이것이 우리가 회귀 모델의 손실 함수로 MSE를 사용하는 강력한 이론적 근거입니다.

2. PMF vs. PDF (핵심 구분)

- **PMF (확률 질량 함수):** 이산형 (셀 수 있음, 예: 방 개수).
 - $P(X = 3) \rightarrow$ ”방 3개 일 확률”. 확률(Mass)을 가짐.
 - (시각화: 막대 그래프)
- **PDF (확률 밀도 함수):** 연속형 (셀 수 없음, 예: 집 면적).
 - $P(X = 1500.0) = 0 \rightarrow$ ”정확히 1500.0 sqft 일 확률은 0”.
 - 확률은 항상 구간(면적)으로 계산됨. (예: $P(1500 < X < 1501)$)
 - (시각화: 부드러운 곡선)

3. 중심극한정리 (CLT)

왜 정규 분포(종 모양)가 중요한가?

모집단이 주사위처럼 생긴 분포(Uniform) 이든, 빼딱한 분포(Skewed) 이든 상관없이, 거기서 뽑은 샘플의 평균(\bar{X})은 n 이 충분히 크다면 무조건 정규 분포를 따릅니다.

4. 다중공선성 (Collinearity) 함정

”침실 수(beds)의 계수가 음수(-89.9)가 나왔습니다. 오류인가요?”

아닙니다. 이는 beds와 sqft가 강하게 연관(collinear)되어 있기 때문입니다. 해석: ”다른 변수, 특히 총 면적(sqft)을 고정한 채로 침실 수만 1개 늘리면 (즉, 방을 억지로 조개면) 집값이 \$89.9k 하락한다”는 합리적인 의미입니다.

5. 이분산성 (Heteroscedasticity)과 추론

”모델의 신뢰 구간(CI)은 어떤 방법을 믿어야 하나요?”

- 공식 기반 CI (t-검정): $[0.544, 0.636]$ (좁음 \rightarrow 지나치게 낙관적)
- 부트스트랩 CI (Resampling): $[0.487, 0.705]$ (넓음 \rightarrow 더 정직함)

결론: 우리 데이터는 ’이분산성’ (집이 클수록 오차가 커짐)을 보였고, 이는 공식 기반 방법의 가정을 위배합니다. 따라서 가정이 깨졌을 때는 부트스트래핑이 더 신뢰할 수 있는 불확실성 추정 방법입니다.