

December 10, 2025

- 강의명: CS109A: 데이터 과학 입문
- 주차: Lecture 08
- 교수명: Pavlos Protopapas, Kevin Rader, Chris Gumb
- 목적: Lecture 08의 핵심 개념 학습

□ 핵심 요약

이 문서는 선형 회귀 모델의 단순한 예측을 넘어, 우리가 얻은 모델이 얼마나 신뢰할 수 있는지(정확성), 모델에 포함된 변수들이 실제로 의미가 있는지(유의성), 그리고 모델의 예측이 얼마나 확실한지(예측 구간)를 평가하는 통계적 추론 방법을 다룹니다.

데이터에 존재하는 불확실성을 이해하고, '부트스트래핑'이라는 시뮬레이션 기법을 통해 회귀 계수($\hat{\beta}$)의 분포를 추정합니다. 이를 바탕으로 신뢰구간을 계산하고, t -검정 및 p -값을 이용해 각 예측 변수의 중요도와 통계적 유의성을 검증하는 방법을 배웁니다. 마지막으로 모델의 '평균 예측'에 대한 신뢰구간과 '개별 예측'에 대한 예측구간의 차이점을 명확히 구분합니다.

Contents

1	개요: 왜 '추론'이 필요한가?	2
2	핵심 용어 정리	3
3	1부: 추정치의 정확성 평가 (Accuracy of Estimates)	4
3.1	문제 제기: 불확실성은 어디에서 오는가?	4
3.2	부트스트래핑 (Bootstrapping): '평행 우주' 시뮬레이션	4
3.3	신뢰구간: β 의 분포에서 정확성 찾기	4
3.3.1	신뢰구간 (Confidence Interval, CI) 계산법	5
3.3.2	표준 오차 (Standard Error, SE)	5
4	2부: 예측 변수의 유의성 평가 (Significance)	6
4.1	무엇이 '중요한' 예측 변수인가?	6
4.2	t -검정 통계량: 신뢰도를 반영한 중요도	6
4.3	p -값: '가장 중요한' 것이 '유의미'한가?	6
4.4	p -값의 정의와 해석	7

5 3부: 예측의 불확실성 (Uncertainty in Predictions)	8
5.1 함수 f 의 신뢰구간 (CI for f)	8
5.2 새로운 y 의 예측구간 (PI for y)	8
5.3 구간의 '나팔' 모양	8
6 빠르게 훑어보기 (1-Page Summary)	10

1 개요: 왜 '추론'이 필요한가?

선형 회귀 모델을 학습시키면 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 와 같은 식을 얻습니다. 예를 들어, TV 광고 예산(x)과 매출(y) 간의 관계가 $\hat{y} = 1.01x + 0.005$ 라고 가정해 봅시다.

이 식의 해석은 "TV 광고 예산을 \$1,000 늘리면 매출이 \$1,010 증가한다 (순이익 \$10)" 입니다. 하지만 이 '1.01'이라는 숫자는 우리가 가진 *하나의* 데이터 샘플로부터 얻은 *추정치* ($\hat{\beta}_1$) 일 뿐입니다.

만약 우리가 다른 날짜에 데이터를 다시 수집했다면(다른 '실현' or 'realization') 어땠을까요? 아마도 $\hat{\beta}_1 = 1.03$ 이나 $\hat{\beta}_1 = 0.98$ 처럼 미묘하게 다른 값을 얻었을 것입니다.

통계적 추론(Inference)은 이러한 불확실성을 다루는 학문입니다. 이 문서의 목표는 다음과 같은 질문에 답하는 것입니다.

- **정확성 (Accuracy):** 우리가 얻은 $\hat{\beta}_1 = 1.01$ 이라는 값은 얼마나 정확하고 신뢰할 수 있는가? (1부)
- **유의성 (Significance):** $\hat{\beta}_1$ 이 '0'과 충분히 멀리 떨어져 있는가? 즉, TV 광고(x)가 매출(y)에 *정말로* 영향을 미치는가, 아니면 그냥 우연인가? (2부)
- **예측 불확실성 (Prediction Uncertainty):** 모델이 예측한 값 \hat{y} 는 얼마나 믿을 수 있는가? (3부)

2 핵심 용어 정리

본격적인 학습에 앞서 주요 용어들을 정리합니다.

Table 1: 선형 회귀 추론 핵심 용어

용어	원어 (English)	쉬운 설명
추론	Inference	제한된 데이터(샘플)를 가지고 더 큰 모집단의 특성을 추측하는 과정
$\hat{\beta}$ (계수 추정치)	Coefficient Estimate	우리가 가진 데이터로 계산한 회귀 계수. (e.g., 1.01)
부트스트래핑	Bootstrapping	원본 데이터에서 '복원 추출'을 반복하여 가상의 데이터셋들을 만드는 기법
복원 추출	Sampling with Replacement	데이터를 뽑은 후 다시 집어넣고 다음 데이터를 뽑는 방식 (중복 허용)
신뢰구간	Confidence Interval (CI)	실제 모수(e.g., 진짜 β_1)가 포함될 것이라 95% 신뢰하는 범위
표준 오차	Standard Error (SE)	$\hat{\beta}$ 값들이 평균으로부터 얼마나 흩어져 있는지를 나타내는 표준편차
가설 검정	Hypothesis Testing	"효과가 없다"(H_0)는 주장이 맞는지 데이터로 검증하는 절차
귀무가설 (H_0)	Null Hypothesis	"아무런 효과가 없다"는 기본 가정. (e.g., " $\beta_1 = 0$ 이다.")
t-검정 통계량	t-test statistic	계수 값이 0으로부터 표준 오차의 몇 배만큼 떨어져 있는지(신호 대 잡음비)
p-값	p-value	귀무가설(H_0)이 맞다고 할 때, 현재 데이터(혹은 더 극단적인)가 *우연히* 관찰될 확률. (작을수록 H_0 가 틀렸다고 확신)
예측구간	Prediction Interval (PI)	*새로운* 데이터 포인트 y 하나가 존재할 것이라 95% 신뢰하는 범위

3 1부: 추정치의 정확성 평가 (Accuracy of Estimates)

3.1 문제 제기: 불확실성은 어디에서 오는가?

우리가 가진 데이터는 완벽하지 않습니다. 불확실성(오차)의 원인은 크게 두 가지입니다.

1. 우연적 오차 (Aleatoric / Irreducible Error, ϵ)

- 시스템에 본질적으로 내재된 무작위성 또는 노이즈입니다.
- 예: 동일한 광고비를 써도 날씨, 경쟁사 프로모션 등 '측정되지 않은' 요인 때문에 매출이 매번 다르게 나옵니다.
- 이는 모델을 아무리 개선해도 줄일 수 없는 '축소 불가능한 오차'입니다.

2. 인식론적 오차 (Epistemic / Misspecification Error)

- 우리가 모델을 잘못 설정했거나(e.g., 비선형인데 선형으로 가정) 데이터가 부족해서 발생하는 오차입니다.
- 우리가 가진 데이터는 수많은 가능한 현실 중 하나의 '실현(realization)'일 뿐입니다.

이러한 오차 때문에, 우리가 데이터를 다시 수집하면 $\hat{\beta}$ 값도 계속 바뀔 것입니다. 우리의 목표는 이 $\hat{\beta}$ 가 얼마나 변동(Variability)하는지 파악하는 것입니다.

3.2 부트스트래핑 (Bootstrapping): '평행 우주' 시뮬레이션

$\hat{\beta}$ 의 변동성을 알려면 "평행 우주"에서 데이터를 여러 번 가져와 $\hat{\beta}$ 를 여러 번 계산해 보면 됩니다. 하지만 현실에선 불가능합니다.

해결책: 부트스트래핑 (Bootstrapping) 우리가 가진 원본 데이터셋(크기 N)을 '모집단' 그 자체라고 간주하고, 이로부터 가상의 '평행 우주' 데이터셋들을 생성하는 기법입니다.

□ 핵심 정보

직관적 비유: 주머니 속 공 뽑기

1. 우리에게 $N = 5$ 개의 공(데이터)이 담긴 주머니(원본 데이터셋)가 있습니다. (공 번호: 1, 3, 5, 8, 9)
2. 이 주머니에서 공을 하나 꺼내 번호를 확인하고(e.g., 8번), 복제본을 만든 뒤, 새 주머니로 옮깁니다.
3. **중요:** 꺼냈던 8번 공은 다시 원본 주머니에 집어넣습니다. (이것이 '복원 추출'입니다.)
4. 다시 주머니에서 공을 꺼냅니다. 아까 뽑았던 8번이 또 나올 수도 있습니다. (e.g., 8번) 복제본을 새 주머니로 옮깁니다.
5. 이 과정을 원본 크기 $N = 5$ 가 될 때까지 반복합니다.
6. **결과:**
 - 원본 샘플: {1, 3, 5, 8, 9}
 - 첫 번째 부트스트랩 샘플: {8, 8, 3, 5, 1} (8번 중복, 9번 누락)
7. 이 16의 과정을 S 번(e.g., 1000번) 반복하여 S 개의 '부트스트랩 샘플'(평행 우주)을 만듭니다.

3.3 신뢰구간: β 의 분포에서 정확성 찾기

부트스트래핑으로 S 개의 '가상' 데이터셋을 만들었습니다. 이제 $\hat{\beta}$ 의 분포를 찾을 수 있습니다.

절차:

1. S 개의 각 부트스트랩 샘플에 대해 선형 회귀 모델을 학습시킵니다.
2. S 개의 서로 다른 $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \dots, \hat{\beta}^{(S)}$ 값들을 얻습니다.

3. 이 값들을 모아 히스토그램을 그리면 $\hat{\beta}$ 의 (근사적인) 샘플링 분포가 됩니다.

이 분포가 바로 $\hat{\beta}$ 의 불확실성을 시각적으로 보여줍니다. 분포가 좁으면(표준편차가 작으면) 추정치가 매우 정확한 것이고, 넓으면(표준편차가 크면) 부정확한 것입니다.

[참고 이미지: $\hat{\beta}_1$ 의 부트스트랩 분포 - TV 광고는 변동성이 작고, Radio 광고는 변동성이 큼]

3.3.1 신뢰구간 (Confidence Interval, CI) 계산법

이 분포를 사용해 "실제 β 값이 존재할 것이라 95% 신뢰하는 구간"을 계산할 수 있습니다.

방법 (가정 없는 백분위수 방식):

1. S 개의 $\hat{\beta}$ 값들을 크기순으로 정렬합니다. (e.g., $S = 1000$ 개)
2. 95% 신뢰구간을 원한다면, 하위 2.5%와 상위 2.5%를 잘라냅니다.
3. 즉, 정렬된 값 중에서 [2.5 백분위수]와 [97.5 백분위수] 값을 찾습니다.
4. (e.g., 1000개 샘플 중 25번째 값과 975번째 값)
5. 예시: [11.50, 12.26, 12.81, ..., 15.21]
 - 'np.percentile(betas, 2.5)' → 12.80 (하한)
 - 'np.percentile(betas, 97.5)' → 13.71 (상한)
 - 95% CI = [12.80, 13.71]

해석: 우리의 $\hat{\beta}$ 추정치가 이 과정을 통해 얻어졌을 때, 실제 (참) β 값이 [12.80, 13.71] 구간 내에 존재한다고 95% 신뢰할 수 있습니다.

3.3.2 표준 오차 (Standard Error, SE)

신뢰구간 외에 불확실성을 요약하는 또 다른 값입니다.

- 표준 오차 (SE): S 개 $\hat{\beta}$ 값들의 표준편차입니다. ($\sigma_{\hat{\beta}}$)
- 근사적 신뢰구간: 만약 $\hat{\beta}$ 의 분포가 정규분포(종 모양)를 따른다고 *가정*한다면, 95% CI는 대략 다음과 같이 근사할 수 있습니다.

$$95\% \text{ CI} \approx [\mu_{\hat{\beta}} - 2 \times SE_{\hat{\beta}}, \mu_{\hat{\beta}} + 2 \times SE_{\hat{\beta}}]$$

(여기서 $\mu_{\hat{\beta}}$ 는 S 개 $\hat{\beta}$ 값들의 평균입니다.)

주의사항

표준 편차(Standard Deviation) vs. 표준 오차(Standard Error)

- 표준 편차 (SD): 데이터 포인트(y)가 평균(\bar{y})으로부터 얼마나 흩어져 있는가? (데이터 자체의 변동성)
- 표준 오차 (SE): 추정치($\hat{\beta}$)가 실제 모수(β)로부터 얼마나 흩어져 있을 것으로 *예상*되는가? (추정치들의 변동성)

부트스트래핑에서는 $\hat{\beta}$ 샘플들의 표준편차를 표준 오차(SE)로 사용합니다.

4 2부: 예측 변수의 유의성 평가 (Significance)

4.1 무엇이 '중요한' 예측 변수인가?

이제 우리는 각 예측 변수(e.g., TV, Radio, Newspaper)의 $\hat{\beta}$ 분포를 알고 있습니다. 어떤 변수가 결과(매출)에 가장 큰 영향을 미칠까요?

- 단순한 방법: $\hat{\beta}$ 의 평균값 $|\mu_{\hat{\beta}}|$ 이 가장 큰 변수.
- 문제점: Figure ??에서 보듯이, Radio 광고($\mu_{\beta_1} = -0.05, \sigma_{\beta_1} = 0.1$)는 평균은 0에 가깝지만 변동성이 매우 큽니다. 이는 실제 값이 0.15일 수도, -0.25일 수도 있음을 의미합니다 (신뢰할 수 없음).
- 반면 TV 광고($\mu_{\beta_1} = 0.05, \sigma_{\beta_1} = 0.005$)는 값은 작지만 변동성이 매우 적어, 0이 아니라고 확실히 말할 수 있습니다.

즉, '중요도'는 계수의 크기(신호)뿐만 아니라 불확실성(잡음)도 함께 고려해야 합니다.

4.2 \hat{t} -검정 통계량: 신뢰도를 반영한 중요도

이 '신호 대 잡음비'를 측정하는 지표가 바로 \hat{t} -검정 통계량입니다. (강의에서는 \sqrt{n} 을 생략한 \hat{t} -test hat 을 사용)

$$\hat{t}\text{-test} = \frac{\mu_{\hat{\beta}}}{\sigma_{\hat{\beta}}} = \frac{\text{계수 평균 (신호)}}{\text{표준 오차 (잡음)}}$$

이 값은 "계수 평균이 0으로부터 표준 오차의 몇 배만큼 떨어져 있는가?"를 의미합니다. \hat{t} 통계량의 절댓값이 클수록, 그 변수는 불확실성에 비해 더 강력한 신호를 가집니다.

□ 핵심 요약

특징 중요도 순위의 변화

캘리포니아 주택 가격 데이터 예시에서, 중요도 순위는 어떤 지표를 쓰느냐에 따라 달라집니다.

- '평균 방 수 (AveBedrms)'는 계수 자체는 크지만, 부트스트래핑 결과 불확실성 (SE) 이 매우 커서 \hat{t} -test 순위는 낮아졌습니다.
- 반면 '중간 소득 (MedInc)'은 계수 값도 크고 불확실성도 낮아, 가장 신뢰할 수 있는 중요한 예측 변수로 선정되었습니다.

4.3 p-값: '가장 중요한' 것이 '유의미'한가?

\hat{t} -test로 '가장 중요한' 변수 (MedInc)를 찾았습니다. 하지만 이런 질문이 남습니다.

"이 변수들이 모두 쓰레기 (junk) 이고, MedInc는 단지 '쓰레기 중 으뜸'인 것은 아닐까?"

즉, MedInc가 매출에 미치는 영향이 실제로 존재하는지, 아니면 우리가 가진 데이터에서 우연히 그렇게 보인 것인지 검증해야 합니다. 이것이 가설 검정 (Hypothesis Testing) 입니다.

- 귀무가설 (H_0): "이 예측 변수는 y 에 아무런 영향을 미치지 않는다."
 - (수학적 표현: 실제 $\beta = 0$ 이다.)
 - (이 경우, 우리가 관찰한 \hat{t} -test 값은 순전히 우연 (random chance)의 산물이다.)
- 대립가설 (H_1): "이 예측 변수는 y 에 유의미한 영향을 미친다."

- (수학적 표현: $\beta \neq 0$ 이다.)

4.4 p-값의 정의와 해석

가설 검증은 H_0 가 맞다고 가정하고 시작합니다.

검증 절차:

1. H_0 가 맞다고 가정합니다. (즉, x 와 y 는 아무 관계가 없다. $\beta = 0$ 이다.)
2. 이 가정 하에서, 데이터를 랜덤 노이즈로 간주하고 \hat{t} -test 값을 계산합니다.
3. 이 과정을 수없이 반복하면, '순전히 우연'에 의해 발생하는 \hat{t} -test 값들의 분포 (Null Distribution) 를 얻을 수 있습니다. (이 분포는 **Student's t-distribution**으로 알려져 있습니다.)
4. **핵심 질문:** 이 '우연의 분포'에서, 우리가 실제 데이터로 계산한 \hat{t} -test 값 (e.g., $t^* = 49$) 또는 그보다 더 극단적인 값이 관찰될 확률은 얼마인가?
5. 이 확률이 바로 **p-값 (p-value)**입니다.

$$p\text{-value} = P(|t_{\text{random}}| \geq |t_{\text{our_model}}^*| \mid H_0 \text{ is true})$$

[참고 이미지: p-값의 시각적 이해 - t-분포 곡선에서 극단적인 꼬리 영역의 면적]

p-값 해석:

- **p-값이 크다 (e.g., $p = 0.5$):**
 - "우리가 관찰한 t^* 값은 '아무 효과가 없다'고 가정해도 50%의 확률로 우연히 발생할 수 있다."
 - $\rightarrow H_0$ 를 기각할 근거가 부족하다. (변수가 유의미하다고 말할 수 없다.)
- **p-값이 작다 (e.g., $p = 0.01$):**
 - "우리가 관찰한 t^* 값은 '아무 효과가 없다'고 가정하면 단 1%의 확률로만 우연히 발생할 수 있다. (매우 희귀한 일)"
 - \rightarrow "우연이라기엔 너무 극단적이다. H_0 가정이 틀렸을 것이다."
 - $\rightarrow H_0$ 를 기각하고 H_1 을 채택한다. (변수가 통계적으로 유의미하다고 결론)

유의 수준 (Significance Level): 통상적으로 $p < 0.05$ (5%)를 H_0 기각의 기준으로 삼습니다.

5 3부: 예측의 불확실성 (Uncertainty in Predictions)

지금까지 β 계수 자체의 불확실성을 다뤘습니다. 이제 모델의 예측값 \hat{y} 의 불확실성을 다룹니다.

5.1 함수 f 의 신뢰구간 (CI for f)

부트스트래핑을 통해 S 개의 다른 모델 $\hat{f}^{(i)}(x) = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)}x$ 을 얻었습니다. 이 S 개의 회귀선을 모두 그리면 '스파게티 플롯(spaghetti plot)'이 됩니다.

[참고 이미지: 1000개의 부트스트랩 모델(회귀선)을 그린 '스파게티 플롯']

이 플롯은 $\hat{\beta}$ 의 불확실성이 $\hat{f}(x)$ 예측의 불확실성으로 어떻게 전파되는지 보여줍니다.

- 특정 x 값(e.g., TV Budget = 200)에서, 1000개의 다른 예측값 $\hat{f}^{(i)}(200)$ 이 나옵니다.
- 이 값들의 95% 백분위수 구간을 계산할 수 있습니다.
- 이 작업을 모든 x 에 대해 수행하여 연결하면, 함수 f 에 대한 95% 신뢰구간(Confidence Interval) 밴드(band)가 만들어집니다.

해석: "우리는 실제 평균 응답을 나타내는 회귀선 $f(x)$ 가 95%의 신뢰로 이 밴드(연한 하늘색 영역) 안에 있다고 믿는다."

5.2 새로운 y 의 예측구간 (PI for y)

$f(x)$ 의 신뢰구간은 모델 자체의 불확실성(즉, $\hat{\beta}$ 가 β 와 다를 수 있는 오차)만 반영합니다.

하지만 우리가 예측하려는 새로운 관측치 y 는 모델 $f(x)$ 에 더해 축소 불가능한 오차 ϵ 까지 포함하고 있습니다. ($y = f(x) + \epsilon$)

[참고 이미지: 신뢰구간(CI)과 예측구간(PI)의 비교 - PI가 ϵ 오차를 추가로 포함하므로 항상 더 넓음]

따라서 '새로운 y 값'이 존재할 범위를 나타내는 예측구간(Prediction Interval)은 $f(x)$ 의 신뢰구간보다 항상 더 넓습니다.

주의사항

신뢰구간 (CI) vs. 예측구간 (PI)

두 개념은 예측 대상이 다릅니다.

- 신뢰구간 (CI): "TV 예산이 \$200,000일 때, 평균 매출($f(x)$)은 95% 신뢰로 [\$16.5M, \$17.5M] 사이에 있을 것이다."
- (오직 $\hat{\beta}$ 의 불확실성만 고려)
- 예측구간 (PI): "TV 예산이 \$200,000인 새로운 매장 하나의 개별 매출(y)은 95% 확률로 [\$14.0M, \$20.0M] 사이에 있을 것이다."
- ($\hat{\beta}$ 의 불확실성 + 개별 오차 ϵ 의 불확실성 모두 고려)

5.3 구간의 '나팔' 모양

Figure ??에서 볼 수 있듯이, 신뢰구간과 예측구간은 데이터의 중심(e.g., x 의 평균 \bar{x})에서 가장 좁고, 양 끝으로 갈수록 넓어지는 '나팔' 또는 '갈매기' 모양을 띕니다.

이유: 회귀선은 데이터의 중심(\bar{x}, \bar{y})을 축으로 회전하는 경향이 있습니다. $\hat{\beta}_1$ (기울기)의 작은 불확실성

이라도, 중심에서 멀어질수록 \hat{y} 예측값에 큰 차이를 만들어내기 때문입니다.

6 빠르게 훑어보기 (1-Page Summary)

□ 핵심 요약

1. 부트스트래핑 (Bootstrapping)

- 목적: $\hat{\beta}$ 추정치의 불확실성(변동성)을 알기 위해.
- 방법: 원본 데이터(크기 N)에서 복원 추출을 N 번 수행하여 '부트스트랩 샘플'을 만든다. 이 과정을 S 번 반복한다.
- 결과: S 개의 모델과 S 개의 $\hat{\beta}$ 분포를 얻는다.

□ 핵심 정보

2. 신뢰구간 (CI) vs. 예측구간 (PI)

- 신뢰구간 (CI): 평균 응답 $\hat{f}(x)$ 의 불확실성.
 - (오차 원인: $\hat{\beta}$ 의 불확실성)
- 예측구간 (PI): 새로운 관측치 y 의 불확실성.
 - (오차 원인: $\hat{\beta}$ 의 불확실성 + 축소 불가능한 오차 ϵ)
 - 항상 PI가 CI보다 넓다.

3. \hat{t} -test 와 p-value (유의성 검증)

- 질문: 이 변수가 y 에 미치는 영향이 우연인가, 실제인가?
- 귀무가설 (H_0): 우연이다 ($\beta = 0$).
- \hat{t} -test 통계량: $\frac{\mu_{\hat{\beta}}}{\sigma_{\hat{\beta}}}$ (신호 대 잡음비). 이 값이 0에서 얼마나 먼가?
- p-value: H_0 가 맞다고 가정할 때, 우리가 관찰한 \hat{t} -test 값(혹은 더 극단적인 값)이 우연히 나올 확률.
- 결론: $p < 0.05$ 이면, "우연히 일어날 확률이 5% 미만이므로 H_0 는 틀렸을 것이다. 이 변수는 유의미하다."