기초통계 과제 보고서: Iris Petal Length (ANOVA & Regression)

2025150101 임예찬 | January 19, 2026

(코드는 ipynb로 제출, 본 문서는 결과 해석 중심)

과제 진행 원칙(명세 준수)

본 과제는 기초적인 통계 검정 절차를 경험하는 것이 목적이므로, 정규성/등분산성 가정이 일부 위배되는 경우가 있더라도 이후 분석은 명세대로 **One-way ANOVA**로 진행하였다. (참고로, 등분산성이 위배될 때는 Welch's ANOVA가 더 견고할 수 있지만, 이번 제출물에서는 비교 언급만 하고 절차는 명세를 따른다.)

1 데이터 및 목표

데이터: Iris 데이터셋 (150 samples, 3 species: setosa/versicolor/virginica)

목표

- (A) 종별 `petal_length` 평균 차이 검정: 시각화 → 가정 점검(정규성, 등분산성) → One-way ANOVA → Tukey HSD 사후검정
- (B) `petal_length` 회귀 예측 모델: 입력(sepal_length, sepal_width, petal_width) → Linear Regression → MSE, R^2, 계수 해석

2 기술통계량(Descriptive statistics)

Species별 `petal_length` 요약 통계는 Table 1와 같다.

Table 1: Species별 Petal length 기술통계량

Species	Mean	Std	Min	Q1	Median	Q3	Max
setosa	1.462	0.174	1.0	1.4	1.50	1.575	1.9
versicolor	4.260	0.470	3.0	4.0	4.35	4.600	5.1
virginica	5.552	0.552	4.5	5.1	5.55	5.875	6.9

3 시각화(Boxplot) 및 해석

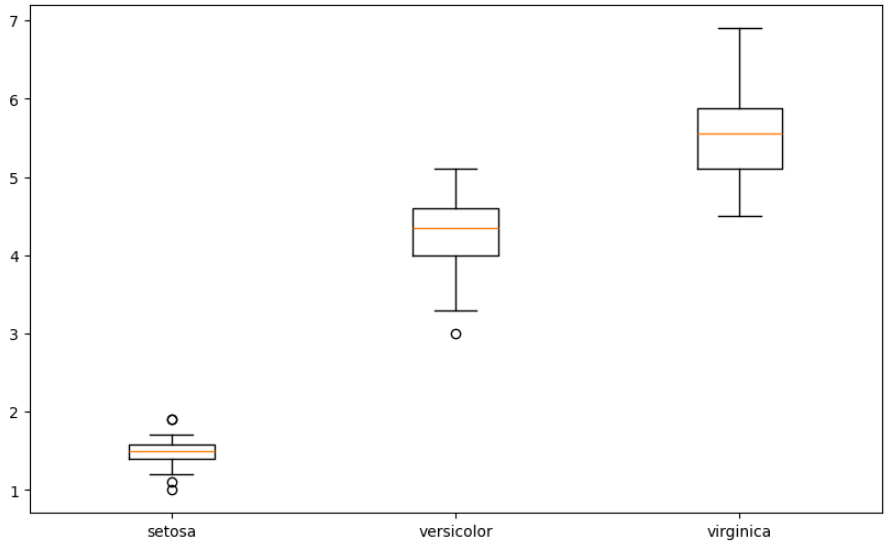

Figure 1: Species별 petal length boxplot

Figure 1에서 중앙값과 분포 범위를 보면 **setosa** < **versicolor** < **virginica** 순으로 꽃잎 길이가 증가한다. 특히 setosa는 다른 두 종과 분포가 거의 겹치지 않아, petal length가 종을 구분하는 강한 특징(feature)임을 시사한다. versicolor와 virginica는 일부 구간에서 분포가 겹치지만, 중앙값과 사분위 범위가 명확히 분리된다.

4 정규성 검정(Shapiro–Wilk)

각 species 그룹에 대해 Shapiro–Wilk 검정을 수행하였다.

가설

$$H_0: \text{해당 그룹 데이터는 정규분포를 따른다,} \qquad H_1: \text{정규분포를 따르지 않는다.}$$

Table 2: Shapiro–Wilk 정규성 검정 결과 ($\alpha = 0.05$)

Species	W statistic	p-value	Decision
setosa	0.9550	0.0548	Fail to reject H_0
versicolor	0.9660	0.1585	Fail to reject H_0
virginica	0.9622	0.1098	Fail to reject H_0

모든 그룹에서 p-value가 0.05보다 커서, 정규성 가정을 기각할 근거가 없다 (Table 2).

5 등분산성 검정(Levene)

Levene 검정으로 등분산성을 점검하였다.

가설

$$H_0: \sigma^2_{\text{setosa}} = \sigma^2_{\text{versicolor}} = \sigma^2_{\text{virginica}}, \qquad H_1: \text{적어도 하나의 분산이 다르다.}$$

$$\text{Levene statistic} = 19.480, \qquad p\text{-value} = 3.129e-08.$$

p-value가 0.05보다 작으므로, 등분산성 가정이 위배됨을 시사한다. 다만 과제 명세에 따라 이후 분석은 One-way ANOVA로 진행하였다.

6 One-way ANOVA

가설

$$H_0: \mu_{\text{setosa}} = \mu_{\text{versicolor}} = \mu_{\text{virginica}}, \qquad H_1: \text{적어도 한 그룹의 평균이 다르다.}$$

$$F = 1180.161, \qquad p\text{-value} = 2.857e-91.$$

p-value가 매우 작으므로 유의수준 0.05에서 H_0를 기각한다. 즉, 세 종의 평균 **petal length**는 동일하다고 보기 어렵고, 적어도 하나의 종이 다른 평균을 가진다.

7 사후검정(Tukey HSD)

ANOVA가 유의하므로 Tukey HSD로 모든 쌍 비교를 수행하였다 (FWER=0.05).

Table 3: Tukey HSD pairwise comparison 결과

Group 1	Group 2	Mean diff	p-adj	Lower	Upper	Reject
setosa	versicolor	2.798	0	2.594	3.002	True
setosa	virginica	4.090	0	3.886	4.294	True
versicolor	virginica	1.292	0	1.088	1.496	True

Table 3에서 모든 쌍 비교가 유의미(**reject=True**)하며, 평균 차이는 **setosa–versicolor**가 약 2.80 cm, **versicolor–virginica**가 약 1.29 cm, **setosa–virginica**가 약 4.09 cm로 나타났다. 따라서 petal length는 세 종을 통계적으로도 명확히 구분한다.

8 결과 요약(ANOVA + Tukey + Boxplot)

- 시각적 결과: boxplot에서 중앙값과 사분위 범위가 setosa < versicolor < virginica 순서로 계단식 분리를 보인다.
- **ANOVA:** 세 그룹 평균이 동일하다는 가설은 기각된다.
- **Tukey HSD:** 모든 종 쌍에서 평균 차이가 유의미하다.

따라서 **setosa**가 가장 짧고, **virginica**가 가장 길며, **versicolor**는 중간이라는 결론을 얻는다.

9 회귀 분석(Linear Regression)

9.1 설정

입력 변수: (sepal_length, sepal_width, petal_width)
타겟 변수: petal_length

Train/Test는 `train_test_split` (random_state=42)로 분리하여 선형회귀(Linear Regression)를 학습하였다.

9.2 성능 지표

$$\text{MSE} = 0.116, \qquad R^2_{\text{train}} = 0.968, \qquad R^2_{\text{test}} = 0.966.$$

R^2가 train/test 모두 약 0.966 수준으로 높아, 선형 모델이 petal length를 상당 부분 설명하고 있음을 보여준다.

9.3 회귀계수 해석

학습된 회귀식은 다음과 같다.

$$\widehat{y} = -0.313 + (0.736) \cdot \text{sepal\_length} + (-0.640) \cdot \text{sepal\_width} + (1.464) \cdot \text{petal\_width}.$$

동일 조건에서 `petal_width`의 계수가 가장 크고 양수이므로, **petal width**가 증가할수록 **petal length**가 증가하는 경향이 가장 강하다. 반면 `sepal_width`는 음의 계수를 가져, 다른 변수가 고정될 때 폭이 커질수록 길이가 감소하는 방향의 관계를 시사한다 (단, 상관/다중공선성 영향이 있을 수 있으므로 단순 인과로 해석하지는 않는다).

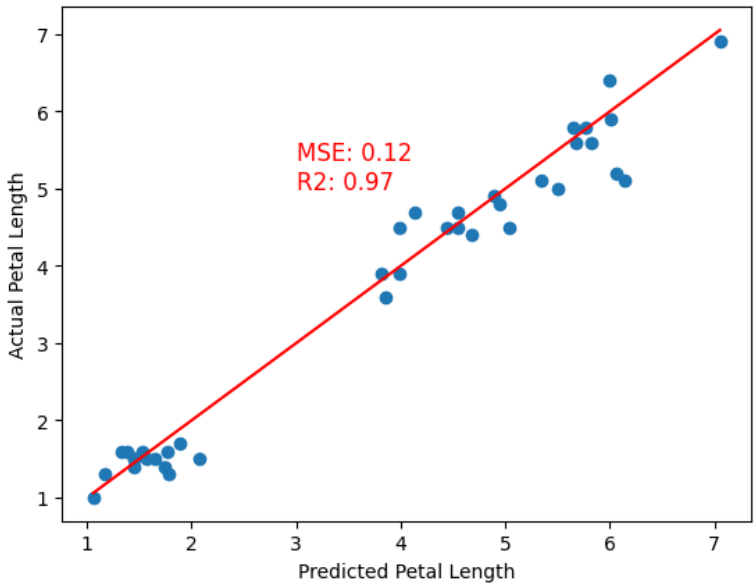

Figure 2: 테스트 세트에서 예측값 vs 실제값 (선형회귀)

Figure 2에서 점들이 대각선(완벽 예측선) 근처에 분포하여 예측이 전반적으로 잘 맞음을 시각적으로 확인할 수 있다.

10 결론

이번 과제는 검정 절차 경험이 목적이므로, 가정 점검 결과(예: 등분산성 위배)를 함께 보고하고, 명세에 따라 One-way ANOVA 및 Tukey HSD를 수행하였다. 추후 심화 분석에서는 등분산성 위배 시 Welch's ANOVA, 또는 변환/비모수 검정도 고려할 수 있다.