IRIS 데이터셋을 사용한 평균 차이 검정과 Petal Length 회귀 예측 모델 구축

YBIGTA 28기 첨단컴퓨팅학부 임수빈

1. 종별 Petal Length 평균 차이 검정

1) 데이터 로드 및 구조 확인

데이터 구조 확인 결과, Iris 데이터셋은 총 150개의 관측치(Row)와 5개의 변수(Column)로 구성되어 있음을 확인했다. 독립변수에 해당하는 sepal_length, sepal_width, petal_length, petal_width는 모두 실수형(float64)이며, 타겟 변수인 species는 문자열(object) 형태이다. 또한 모든 변수에서 결측치(Null 값)가 발견되지 않아 별도의 결측치 처리가 필요하지 않았다.

```
Non-Null Count
--------------
150 non-null
150 non-null
150 non-null
150 non-null
150 non-null
```

[그림 1. 결측치 없음을 보여주는 iris.info() 결과]

2) 기술 통계량

```
**Species별 Petal Length 기술 통계량**
            count   mean       std  min  25%   50%    75%  max
species
setosa       50.0  1.462  0.173664  1.0  1.4  1.50  1.575  1.9
versicolor   50.0  4.260  0.469911  3.0  4.0  4.35  4.600  5.1
virginica    50.0  5.552  0.551895  4.5  5.1  5.55  5.875  6.9
```

[그림 2. 기술 통계량 출력 결과]

Species별 Petal Length의 기술통계량을 분석한 결과, 평균값은 Virginica(5.552) > Versicolor(4.260) > Setosa(1.462) 순으로 나타났다. 표준편차를 볼 때 Setosa(0.174)가 가장 작아 데이터가 평균 근처에 밀집해 있는 반면, Virginica(0.552)는 상대적으로 퍼져 있음을 알 수 있다.

3) 시각화 결과 해석

1. setosa는 petal_length가 다른 두 종에 비해 현저히 짧으며, 분포의 범위도 매우 좁고 일정하다.

2. virginica는 가장 긴 꽃잎 길이를 가진다.
3. versicolor species는 setosa와 virginica 중간에 분포한다.
4. 세 종의 박스(IQR)가 서로 겹치지 않는 것으로 보아, petal_length만으로도 종을 구분하는 데 매우 유용한 변수임을 알 수 있다.

4) 정규성 검정 (Shapiro-Wilk)

(1) 가설 수립

각 종별 정규성 검정을 위한 가설은 다음과 같다.

- setosa에 대한 가설

귀무가설(H0) : setosa 의 데이터는 정규분포를 따른다.
대립가설(H1) : setosa 의 데이터는 정규분포를 따르지 않는다.

- versicolor에 대한 가설

귀무가설(H0) : versicolor 의 데이터는 정규분포를 따른다.
대립가설(H1) : versicolor 의 데이터는 정규분포를 따르지 않는다.

- virginica에 대한 가설

귀무가설(H0) : virginica 의 데이터는 정규분포를 따른다.
대립가설(H1) : virginica 의 데이터는 정규분포를 따르지 않는다.

(2) 정규성 검정 결과 (p-value 해석)

각 종별 정규성 검정 결과는 다음과 같다.

- setosa의 결과

검정통계량: 0.9550, p-value: 0.0548
=> p-value >= 0.05 이므로 정규성을 만족합니다

-versicolor의 결과

검정통계량: 0.9660, p-value: 0.1585
=> p-value >= 0.05 이므로 정규성을 만족합니다.

- virginica의 결과

검정통계량: 0.9622, p-value: 0.1098
=> p-value >= 0.05 이므로 정규성을 만족합니다.

5) 등분산성 검정

(1) 가설 수립

귀무가설(H0): 세 그룹(Species)의 분산은 동일하다.
대립가설(H1): 적어도 한 그룹의 분산은 다르다.

(2) 검정 결과 및 해석

검정통계량은 19.4803이며, p-value값은 0.0000이 나온다.
p-value < 0.05 이므로 귀무가설이 기각되며 등분산성을 만족하지 못한다. 즉, 세 그룹의 분산은 통계적으로 다 다르다고 할 수 있다.

(단, 본 과제의 가이드라인에 따라 등분산성을 만족한다고 가정하고 이후 분석(ANOVA)을 진행한다.)

6) ANOVA 분석

(1) 가설 수립

귀무가설(H0): 세 species 간 Petal Length의 평균은 모두 같다.
대립가설(H1): 적어도 하나의 species는 평균이 다른 종과 다르다.

(2) 분석 결과

F값은 1180.1612이며, p-value값은 2.8568e-91이다. 이 때, p-value값이 매우 작아 지수 표기법으로 나타내었다.

(3) 분석 결과 해석

p-value < 0.05 이므로 귀무가설이 기각된다. 즉, 적어도 한 그룹의 분산이 다르다는 의미이다. 따라서 세 species 간 Petal Length의 평균에는 통계적으로 유의미한 차이가 있다고 할 수 있다.

7) #8. 사후 분석

ANOVA 분석 결과 유의미한 차이가 발견되어 Tukey HSD 사후 검정을 실시하였다. 그 결과 모든 종 간의 비교에서 P-value가 유의 수준 0.05 미만으로 나타나 통계적으로 유의미한 평균 차이가 있음이 확인되었다.

평균 차이(meandiff)를 분석하면 Setosa < Versicolor < Virginica 순으로 Petal Length가 길어지는 경향을 보인다. 구체적으로 Virginica와 Setosa 간의 차이가 약 4.09로 가장 컸으며, Versicolor와 Virginica 간의 차이는 약 1.29로 나타났다.

정리하면 세 가지 종은 Petal Length가 서로 다 다르며, Virginica가 가장 길고 Setosa가 가장 짧다는 것을 알 수 있다.

8) #9. 결과 요약

Boxplot 시각화와 ANOVA 분석 결과, 세 종(Species) 간의 Petal Length 평균에는 뚜렷한 차이가 있었다. 특히 Tukey HSD 사후검정을 통해 Virginica > Versicolor > Setosa 순서로 꽃잎 길이가 통계적으로 유의미하게 길다는 것을 확인할 수 있다. (모든 그룹 간 p-value < 0.05). 따라서 Virginica 종의 꽃잎이 가장 길고, Setosa 종의 꽃잎이 가장 짧다고 결론지을 수 있다.

2. Petal Length 회귀 예측 모델 구축

(1) 회귀 분석 결과값

평균 제곱 오차(MSE)는 0.1300, 결정 계수(R^2 Score)는 0.9603가 나왔다. 회귀 계수(Coefficients)로 sepal_length는 0.7228, sepal_width는 -0.6358, petal_width는 1.4675, 절편 (Intercept)은 -0.2622가 나왔다.

(2) 회귀분석 결과 해석

R^2 점수가 0.96라는 것은 이 모델이 데이터의 변동성을 약 96.0% 설명한다는 뜻으로,매우 높은 예측 성능을 보여준다고 할 수 있다.

변수별로 영향력을 해석해보면, 회귀 계수 중 Petal Width의 계수가 약 1.47로 가장 크게 나타났다. 이는 다른 조건이 동일할 때, 꽃잎 너비(Petal Width)가 1 증가하면 꽃잎 길이(Petal Length)는 약 1.47만큼 증가한다는 양의 상관관계를 의미한다. 따라서, 꽃잎 너비가 꽃잎 길이를 예측하는 데 가장 중요한 영향을 미치는 변수라고 해석할 수 있다.