YBIGTA 3-1 기초통계 레포트 과제

작성자: 28기 남건우

1. 개요

본 분석은 통계학의 기초적인 기법들을 활용하여 Iris(붓꽃) 데이터셋의 특성을 파악하고, 특히 꽃잎의 길이 Petal Length가 종별로 어떤 차이를 보이는지 검증하는 데 목적이 있다. 나아가 꽃받침의 길이와 너비, 꽃잎의 너비 데이터를 활용하여 꽃잎의 길이를 예측하는 회귀 모델을 구축함으로써 데이터 간의 인과관계와 예측 가능성을 탐색하고자 한다.

2. 기술통계 및 EDA

분석에 앞서 전체적인 데이터의 구조를 파악하였다. Iris 데이터셋은 Setosa, Versicolor, Virginica 세 가지 종으로 구성되어 있으며, 각 종별로 꽃잎 길이의 분포를 살펴본 결과는 다음과 같다.

1. Setosa: 평균 약 1.46cm로 가장 짧으며, 데이터의 변동성이 작고 매우 일관된 특징을 보인다.
2. Versicolor: 평균 약 4.26cm로 중간 수준의 길이를 나타낸다.
3. Virginica: 평균 약 5.55cm로 세 종 중 가장 긴 꽃잎을 가졌으며, 분포의 범위가 가장 넓게 나타났다.

시각화를 위해 작성한 Boxplot을 보면, Setosa 종은 다른 두 종과 완전히 분리된 분포를 보여주어 꽃잎 길이만으로도 충분히 분류가 가능함을 시사한다. 반면 Versicolor와 Virginica는 일부 구간이 겹치지만, 중앙값의 위치를 통해 값의 차이가 있음을 확인할 수 있다. 이러한 분포 차이가 통계적으로 유의미한 것인지 확인하기 위해 통계적 가설 검정을 수행하였다.

3. 통계적 가정 검정

ANOVA(분산 분석)를 수행하기 전, 데이터가 통계적 가정을 만족하는지 검토하였다.

1. 정규성 검정 (Shapiro-Wilk Test): 각 그룹별로 정규분포를 따르는지 확인하였다. 모든 그룹에서 유의수준 0.05를 상회하여 정규성을 만족한다고 가정하고 분석을 진행하였다.
2. 등분산성 검정 (Levene's Test): 세 그룹 간의 분산이 동일한지 검정한 결과, p-value가 3.1288e-08로 나타나 귀무가설을 기각하였다. 즉, 그룹 간 분산은 통계적으로 다르다고 판단되나, 본 과제의 지침에 따라 등분산성을 만족한다는 전제하에 ANOVA를 실시하였다.

4. One-way ANOVA

세 가지 종에 따른 꽃잎 길이 평균의 차이가 통계적으로 유의미한지 확인하기 위해 가설을 수립하고 ANOVA를 실시하였다.

1. 귀무가설(Null hypothesis): 세 종의 꽃잎 길이 평균은 모두 같다.
2. 대립가설(Alternative hypothesis): 적어도 한 종의 평균은 다른 종과 다르다.

분석 결과:

- F-statistic: 1180.1612
- p-value: 2.8568e-91

p-value가 유의수준 0.05보다 압도적으로 작으므로 귀무가설을 기각한다. 즉, 세 종의 꽃잎 길이는 통계적으로 매우 유의미한 차이가 있음이 증명되었다.

5. 사후 검정 (Tukey HSD)

ANOVA 결과를 통해 차이가 있음을 확인하였으므로, 구체적으로 어떤 종들 사이에 차이가 있는지 Tukey HSD 검정을 통해 확인하였다. 분석 결과, 모든 종의 쌍(Setosa-Versicolor, Setosa-Virginica, Versicolor-Virginica)에서 Reject=True가 나타났다. 이는 세 종이 서로 모두 통계적으로 유의미하게 다른 꽃잎 길이를 가지고 있음을 의미하며, 길이는 Setosa < Versicolor < Virginica 순으로 길어지는 것을 최종 확인하였다.

6. 회귀 분석을 통한 예측 모델링

마지막으로 꽃잎의 길이 Petal length를 종속 변수로 설정하고, 나머지 세 변수 Sepal length, Sepal width, Petal width를 독립 변수로 하는 선형 회귀 모델을 구축하였다. 앞선 분석이 종에 따른 평균 차이를 검증하는 데 초점을 맞췄다면, 이 파트에서는 데이터셋의 다른 여러 변수가 Petal Length를 얼마나 잘 설명할 수 있는지 확인하고자 회귀 분석을 수행하였다.

- 모델 평가: R square(결정계수)는 0.9603으로 나타났으며, 이는 본 모델이 꽃잎 길이 변동의 약 96%를 설명하는 우수한 성능을 가졌음을 보여준다. MSE 또한 0.1300으로 매우 낮게 측정되었다. 이는 Petal Length가 다른 변수들과 매우 강한 선형 관계를 가짐을 시사한다.
- 변수 영향력: 회귀계수 분석 결과, Petal width(1.4675)가 꽃잎 길이에 가장 큰 양(+)의 영향을 미치는 핵심 변수임을 확인하였다.

7. 결과 해석 및 결론

본 분석에서는 Iris 데이터셋을 대상으로 종에 따라 Petal Length가 유의미한 차이를 보이는지를 단계적으로 검증하였다. 기술통계 및 시각화 결과, Setosa, Versicolor, Virginica 간 꽃잎 길이 분포에 뚜렷한 차이가 관찰되었으며, 이를 One-way ANOVA를 통해 통계적으로 검정한 결과 세 종 간 평균 차이는 매우 유의미함을 확인하였다($p < 0.05$).

이어 수행한 Tukey HSD 사후 검정에서는 모든 종의 쌍에서 유의미한 차이가 나타났으며, Petal Length는 Setosa < Versicolor < Virginica 순으로 증가하는 경향을 보였다. 이는 꽃잎 길이가 Iris 종을 구분하는 핵심적인 특성 중 하나임을 시사한다.

또한 회귀 분석 결과, Sepal length, Sepal width, Petal width를 활용한 선형 회귀 모델은 Petal Length 변동의 약 96%를 설명하며 높은 예측 성능을 보였다($R^2 = 0.9603$). 특히 Petal width가 가장 큰 영향을 미치는 변수로 나타나, 꽃잎 관련 특성 간의 강한 선형 관계를 확인할 수 있었다.

종합적으로 본 분석은 탐색적 분석 → 가설 검정 → 사후 검정 → 예측 모델링의 흐름을 통해, Petal Length가 종 구분과 예측 모두에서 중요한 역할을 하는 변수임을 통계적으로 해석하고 확인하였다는 점에서 의의를 가진다.