데이터과학기초 2차과제

컴퓨터학부 2017110762 노태현

기본사항

6 종류의 생태학적 범주로 나눌 수 있는 새들의 뼈의 길이와 지름의 크기에 대한 데이터

데이터 출처: https://www.kaggle.com/zhangjuefei/birds-bones-and-living-habits

Attributes: 11개

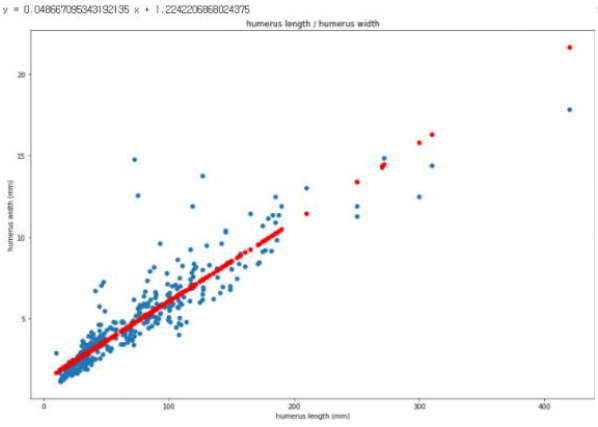
- 생태학적 범주: Swimming Birds(물새, **SW**), Wading Birds(섭금류, **W**), Terrestrial Birds(타조목, **T**), Raptors(맹금류, **R**), Scansorial Birds(딱따구리목, **P**), Singing Birds(참새목, **SO**)
- **humerus**(상완골) length, diameter, **ulna**(자뼈) length, diameter, **femur**(대퇴골) length, diameter, **tibiotarsus**(경골) length, diameter, **tarsometatarsus**(부척골) length, diameter

1차 과제에서 추정한 내용

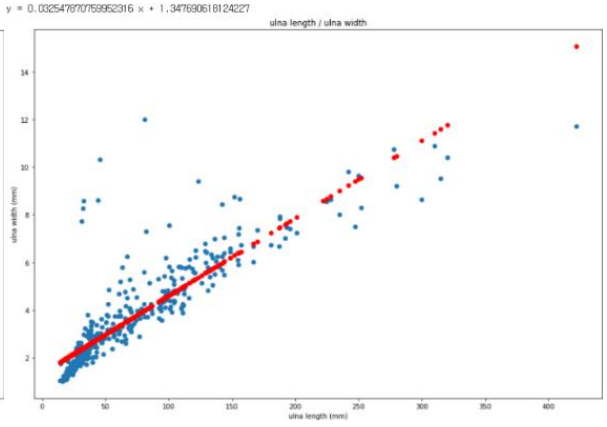
- 뼈의 길이를 고려했을 때, 새의 날개에 있는 두 뼈인 humerus, ulna의 길이 사이에는 서로 상관관계가 있을 것으로 추정된다.
- 뼈의 길이를 고려했을 때, 새의 다리에 있는 세 뼈인 femur, tibiotarsus, tarsometatarsus의 길이 사이에는 서로 상관관계가 있을 것으로 추정된다.
- tarsometatarsus를 제외한 뼈들은 길이가 길어질수록 대부분 지름도 커지는 경향이 보이므로 뼈의 길이와 지름 간에 상관관계가 있다고 할 수 있다.

데이터 분석 1. Linear Regression

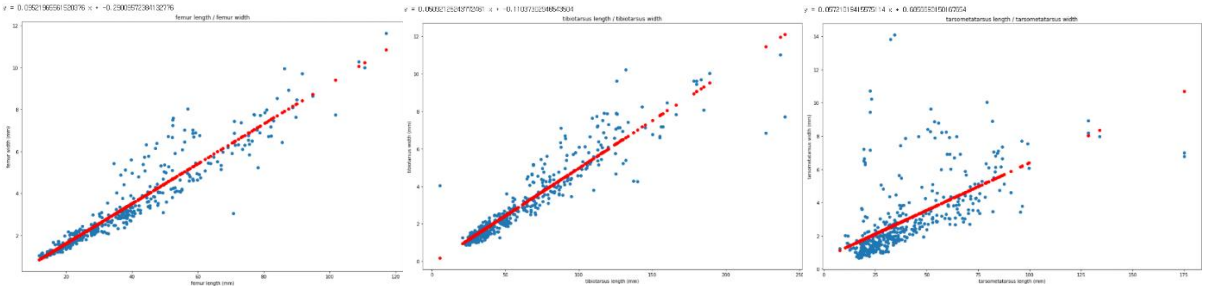
(1) 각 뼈들의 length, width에 대한 수식 계산



length, width 간 변동성: 약 26.1%



약 29.8%

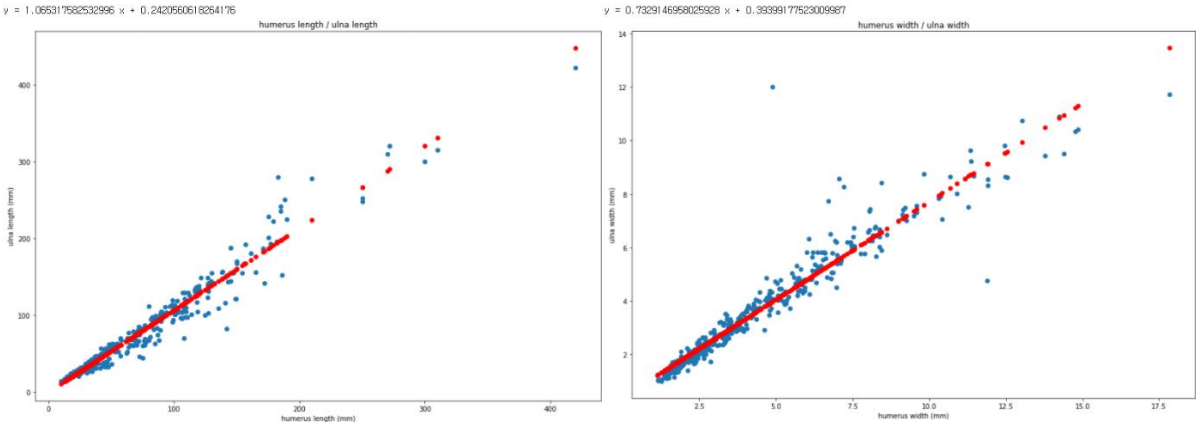


length, width 간 변동성: 약 20.6%　　약 24.2%　　약 59.3%

- humerus, ulna, femur, tibiotarsus의 경우 20%대의 작은 변동성을 가지므로 추정한 것처럼 뼈의 길이가 길어질수록 지름이 커지는 경향을 보인다고 할 수 있음. 따라서 뼈의 길이를 회귀분석을 통해 얻은 수식에 대입했을 때, 20대의 오차율을 가지는 뼈의 지름을 알아낼 수 있다.
- tarsometatarsus의 경우 변동성이 59.3%로 꽤나 크므로, 뼈의 길이가 길어질수록 지름이 커진다고 하기 어려움. 따라서 뼈의 길이를 회귀분석을 통해 얻은 수식에 대입한다고 해도 오차율이 약 59.3% 정도이므로 예측치와 실제 값 사이의 오류가 클 가능성이 높다.

(2) 각 뼈들 간의 length, width에 따른 수식 계산

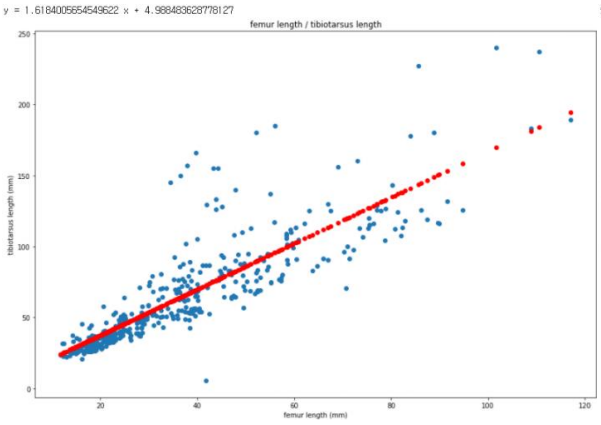
(2-1) 새의 날개에 있는 두 뼈인 humerus, ulna 사이의 상관관계

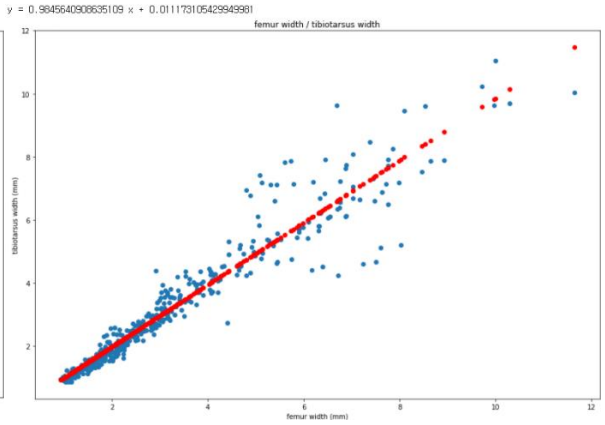


humerus, ulna 간 변동성: length: 약 18.4%　　width: 약 17.5%

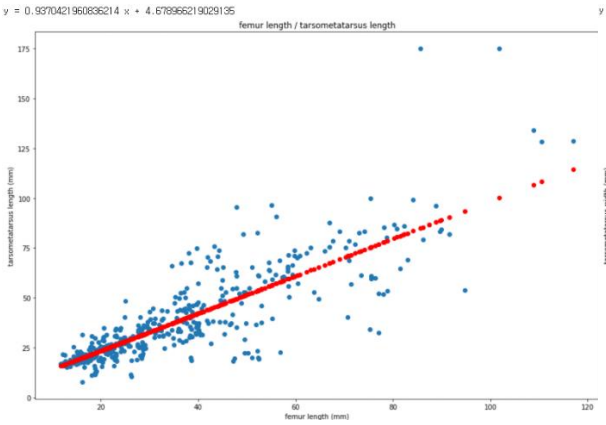
- 대체로 humerus의 길이나 지름이 증가함에 따라 ulna의 길이나 지름 역시 증가하는 경향을 보임. 또한 변동성이 뼈의 길이의 경우 약 18.4%, 지름의 경우 약 17.5%로 작은 변동성을 가지므로, 추정한 것처럼 두 뼈 사이에는 상관관계가 있다고 할 수 있다.
- 따라서, 두 뼈 중 한 뼈의 길이나 지름을 알면, 날개에 있는 다른 뼈의 길이나 지름 역시 회귀분석을 통해 얻은 수식에 대입하면서 예측할 수 있음. 이 때 변동성이 약 18.4%, 17.5%으로 작은 편이므로 예측치와 실제 값 사이의 오차는 작을 것이다.

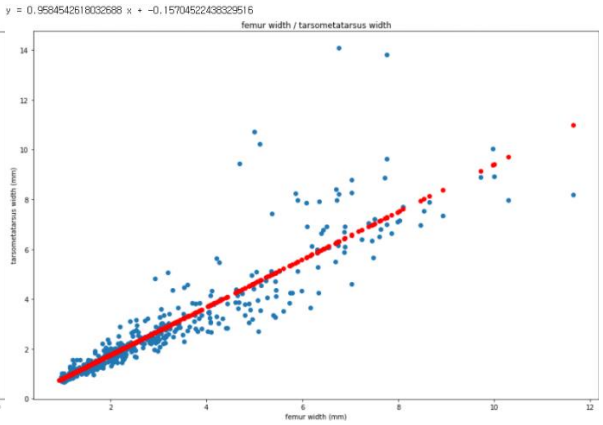
(2-2) 새의 다리에 있는 세 뼈인 femur, tibiotarsus, tarsometatarsus 사이의 상관관계



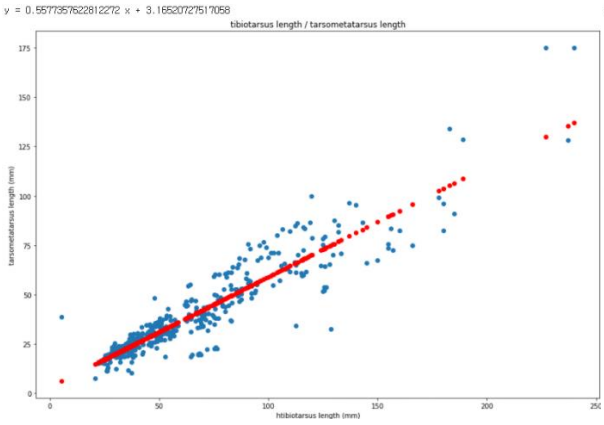


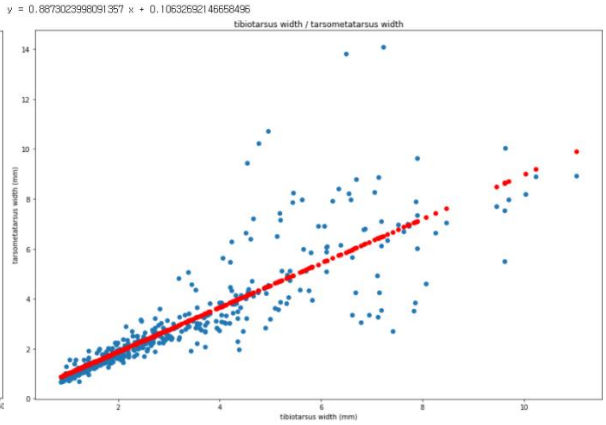
femur, tibiotarsus 간 변동성: length: 약 29.9% width: 약 18.4%





femur, tarsometatarsus 간 변동성: length: 약 32.8% width: 약 32.3%





tibiotarsus, tarsometatarsus 간 변동성: length: 약 22.8% width: 약 38.6%

- 대체로 femur의 길이나 지름이 증가함에 따라 tibiotarsus나 tarsometatarsus의 길이나 지름 역시 증가하는 경향을 보임. 또한 변동성이 뼈의 femur와 tibiotarsus 사이에는 길이의 경우 약 29.9%, 지름의 경우 약 18.4%, femur와 tarsometatarsus 사이에는 길이의 경우 약 32.8%, 지름의 경우 약 32.3%, 마지막으로 tibiotarsus와 tarsometatarsus 사이에는 길이의 경우 약 22.8%, 지름의 경우에는 38.6%의 변동성을 보이므로 세 뼈 사이에는 상관관계가 있다고 할 수 있다.

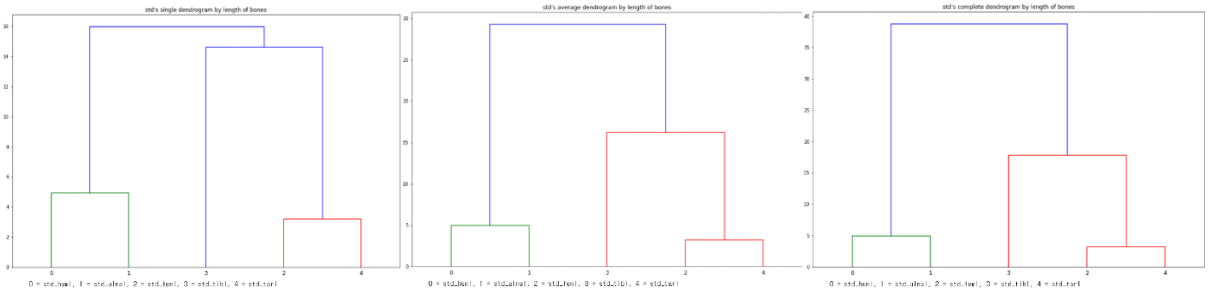
- tarsometatarsus의 지름을 다른 뼈들의 지름과 비교했을 때는 30%가 넘는, 각 뼈들의 길이를 비교했을 때에 비해 비교적 큰 변동성을 보이는데, 이는 앞서 tarsometatarsus의 길이, 지름을 비교했을 때 변동성이 약 59.3%로 상관관계가 크다고 하기 어려운 상황에서 기인한 것이라고 추정된다.

- 새의 다리에 있는 세 뼈들도 길이나 지름을 알면 나머지 뼈들의 길이나 지름 역시 회귀분석을 통해 얻은 수식에 대입하면서 예측할 수 있음. 이 때 변동성에 따라 예측치와 실제 값 사이에 오차가 생기는데, tibiotarsus의 길이를 통해 tarsometatarsus의 길이를 구하는 경우(변동성 약 22.8%)를 제외하면 tarsometatarsus의 길이나 지름을 구하는 데는 변동성이 30%를 넘기 때문에 오차가 길이나 지름을 예측한 값에 비해 커질 것이다.

데이터 분석 2. Hierarchical Clustering

계층적 군집화를 시행하기에 앞서, 데이터 상 humerus length, width, ulna length, width, femur length, width, tibiotarsus length, width, tarsometatarsus length, width가 서로 비슷한 값이 많기에 각 데이터들의 표준편차를 구해서 그 데이터들을 바탕으로 계층적 군집화를 시행하였다.

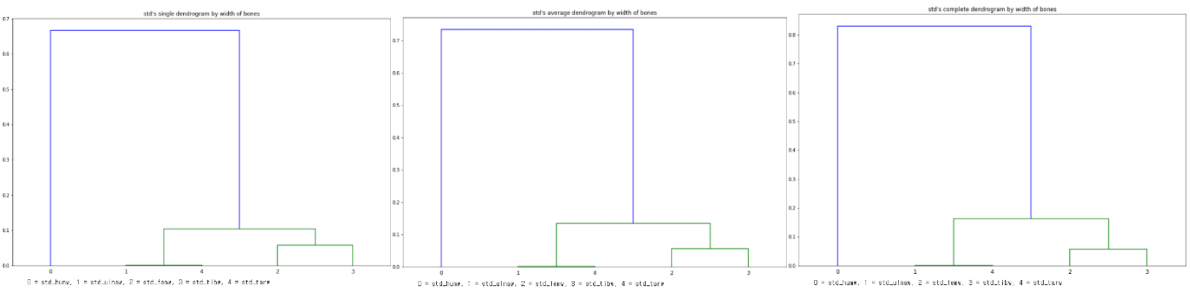
(1) 뼈의 길이에 대한 표준편차들의 계층적 군집화



단일연결법 평균연결법 최장연결법

- 각 연결법을 이용하여 뼈의 길이에 대한 표준편차의 덴드로그램을 그려보았을 때, 클러스터가 2개 정도로 나뉜다고 할 수 있다. 이는 날개의 두 뼈(0: humerus, 1: ulna)와 다리의 세 뼈(2: femur, 3: tibiotarsus, 4: tarsometatarsus)로 뼈의 길이에 대한 표준편차는 다리와 날개로 분리되어 뼈들이 군집화가 된다고 볼 수 있다.

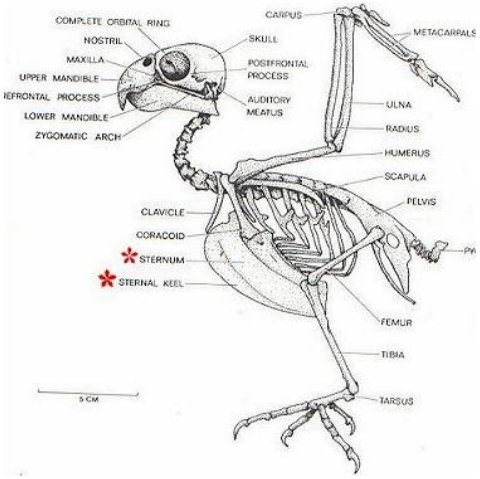
(2) 뼈의 지름에 대한 표준편차들의 계층적 군집화



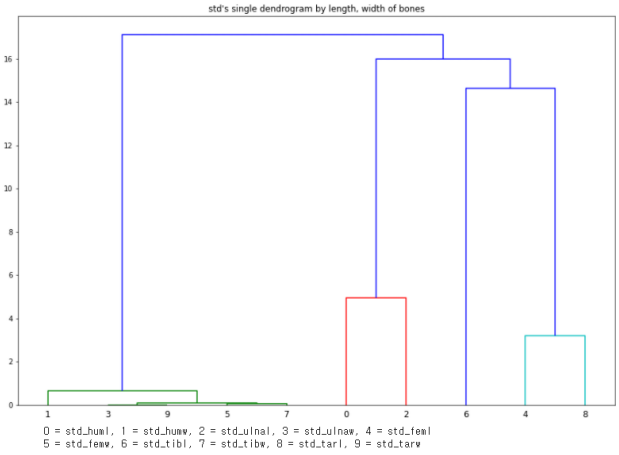
단일연결법 평균연결법 최장연결법

- 각 연결법을 이용하여 뼈의 지름에 대한 표준편차의 덴드로그램을 그려보았을 때, 날개에 있는 뼈인 ulna가 다리에 있는 뼈들과 군집이 같이 묶인다. 이는 이 뼈들의 평균이 대체로 3~4mm, 표준편차가 대체로 2~3mm로 큰 차이가 없기 때문이다.

- 그림에서 보이듯이, humerus는 새의 날개에서 가장 중요한 뼈를 담당하는데, 새가 날기 위해선 humerus의 지름이 큰 영향을 미치며, 이 때문에 뼈의 지름이 다리나 날개의 다른 뼈들에 비해 가장 커질 수밖에 없으므로, 표준편차들을 이용해 뼈의 지름에 대한 군집화를 시행하면 humerus를 제외한 뼈들과 humerus로 군집이 나뉘어진다.



(3) 뼈의 지름, 길이에 대한 표준편차들의 계층적 군집화



단일연결법

- 길이, 지름의 표준편차들을 이용해 덴드로그램을 그려보았을 때, 모든 뼈의 지름에 대한 표준편차 값들은 함께 묶여 클러스터를 형성하고, 뼈의 길이에 대한 표준편차 값들도 클러스터를 형성하는데, 여기서 새의 날개에 있는 두 뼈들과 새의 다리에 있는 세 뼈들끼리 클러스터로 분리가 된다.

- 뼈의 지름에 대한 표준편차 값들의 차이는 매우 작지만, 뼈의 길이에 대한 표준편차 값들의 클러스터와 분리되어 뼈의 지름의 표준편차 값들끼리 함께 묶여 있다.

- 뼈의 길이의 표준편차에 대한 클러스터는 다리와 날개의 뼈들로 분리할 수 있는데, 여기서 다리의 세 뼈들은 tibiotarsus와 그 외의 뼈들로 다시 분리가 가능하다. 이는 대체로 새들의 tibiotarsus가 다리의 다른 두 뼈들에 비해 훨씬 길기 때문이다.

- 따라서 표준편차를 이용해 군집화를 시행할 경우, 각 뼈들의 지름, 날개의 뼈들의 길이, 그리고 다리의 뼈들의 길이로 세 개의 클러스터로 군집화가 된다고 볼 수 있다.

결론

- 회귀를 통해 뼈의 길이와 지름에 대한 비례관계 및 수식을 얻을 수 있고, 날개의 두 뼈 간의 상관관계, 다리의 세 뼈 간의 상관관계 및 수식을 추정한 것처럼 얻을 수 있다. 또한 tarsometatarsus의 경우에는 길이가 길어진다고 지름이 무조건적으로 증가한다고 하기 힘들다고 할 수 있다.
- 군집화를 통해 뼈의 너비와 뼈의 길이에 대한 표준편차는 따로 군집을 형성하며, 뼈의 길이는 날개의 뼈들과 다리의 뼈들로 분리할 수 있는데, 이는 일반적으로 날개의 뼈들과 다리의 뼈들이 각각의 군집을 형성한다고 볼 수 있다.
- 종합해서, 1차 과제에서 추정한 것처럼 tarsometatarsus를 제외한 뼈들은 대체로 길이가 길어질수록 지름이 커진다고 할 수 있고, 날개에 있는 두 뼈들은 서로 상관관계가 있고 뼈의 길이에 대해 그룹화할 수 있으며, 다리에 있는 세 뼈들도 서로 상관관계가 있으며 뼈의 길이에 대해 그룹화할 수 있다.