

[식용 작물의 성장 예측

- 다중선형분석을 통한 변수 확인 및 성장 예측]



**■ 과 목 명 : 시계열 자료 분석과 예측**

**■ 담당교수 : 김명석 교수님**

**■ 제 출 일 : 2022.11.06**

**■ 학과: 데이터 사이언스. 인공지능**

**■ 학번: A64037**

**■ 성명: 서승덕**

**분석의 동기(problem motivation)**

최근 러시아, 우크라이나 전쟁으로 인한 최대 밀 생산국 중 하나인 우크라이나 밀 생산량 감소하였으며, 기후 변화(가뭄)으로 미국, 프랑스 등의 국가에서 밀 생산량 감소하는 등 전쟁, 기후 변화, 수출 제한 등 여러 악재가 맞물리며 식량 가격이 큰 폭으로 상승하였다.

그에 따라, 식용작물의 생산이 중요해지고 있는데, Iot의 발달로 충분히 생육환경을 실시간으로 제어할 수 있는 여건이 마련이 되었고, 그러한 기술을 바탕으로 스마트 팜을 만들어 생육환경 제어한다면 높은 생산성과 품질 향상을 이루어 낼 수 있을 것이라고 판단하였다.

따라서, 스마트팜 제어의 기초가 될 생산 품질 및 생산량에 영향을 주는 생육환경조성 모델을 만듬으로서 최적의 생산조건을 탐색하고자 이 연구를 진행하게 되었다.

**분석의 목적(research objective)**

최적의 성장 예측 모델로 환경인자를 조절한다면 안정적이고 최선의 품질을 지닌 식용작물의 생산이 가능할 것으로 기대 됨에 따라, 영향을 줄 수 있는 각 변수를 탐색하여 어떠한 변수가 영향역이 크고, 양의 관계 및 음의 관계를 밝혀냄으로써 효율적인 생육환경조성 모델을 만들어 보고자 한다.

**문헌연구(literature review)**

스마트팜 데이터를 이용한 토마토 최적인자에 관한 연구\*는 토마토 수확량에 주요한 환경인자를 구하고, 이를 이용한 데이터 분석 방법을 설명하는 연구이며, 스마트팜 토마토 농가의 실제 수집된 자료 활용하였고, 환경인자에 대해서 적절한 변수 변환 후 새롭게 생성된 변수들을 이용하여 모델링하였다. 수확량과 환경변수의 연관성 확인 및 수확량을 예측하기 위한 다중회귀분석을 사용하였고, 회귀계수를 통하여 검정 결과 확인하였으며, 통계적 모형을 이용하여 토마토의 수확량에 영향을 미치는 최적환경인자 확인하고, 토마토 농가의 수확량 예측하였다. 단계적 선택법에 의해 수확 1주전 평균온도, 7주전 일사량, 5주전 물 흡수량, 7주전 공급 PH농도를 변수로 선정하였다.

머신러닝 기반 식용 작물별 생산량 예측 연구\*는 머신러닝과 딥러닝 기법을 접목하여 대한민국에서 생산되는 다양한 식용 작물들의 생산량을 재배 면적과 기후 데이터를 활용하여 예측하는 연구를 진행하였으며, 분리 학습법을 통한 모델링으로 최적의 모델 확인하고, 다양한 방법들로 모델링 하였을 때 가장 성능이 좋은 생산량 예측 모델을 확인하는 구조로 소개하고 있다. 변수 중 일사량, 평균 풍속과 소형 증발량 변수가 가장 많은 작물에 음의 영향이 있음을 확인하였고, 과채류, 조미채소, 과실류는 USD를 통해 구축한 모델의 성능이 적합하였다고 확인하였으며, 과채류와 엽채류를 제외한 작물들은 최근 6 개년 데이터만을 활용한 모델링이 더 좋은 성능을 나타냄을 확인할 수 있었다.

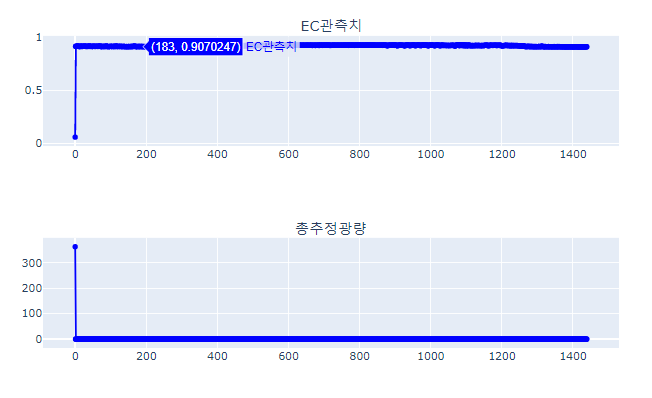
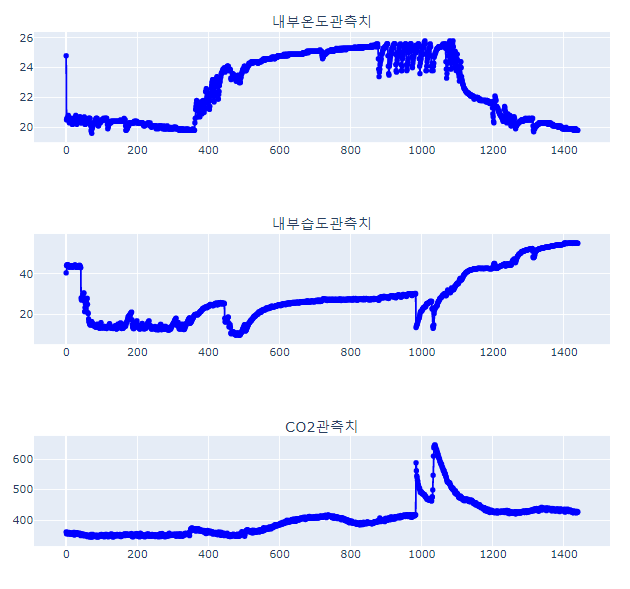
**자료 및 방법론 설명(data and method explanation)**

변수 설명:

* 종속변수
  + 청경채 잎 면적 증감률(label).
* 독립변수
  + 내부 환경 변수(식물성장에 직접적인 영향을 주는 요인): "내부온도관측치", "내부습도관측치", "CO2관측치", "EC관측치"
  + 외부 환경 변수(시설하우스에서 외부 환경 변수는 식물 생장에 큰 영향을 주지 않음.): "외부온도관측치", "외부습도관측치",
  + 설비 변수(설비는 내부 환경을 조절하는 용도로, 설비의 가동 결과가 내부 환경 변수의 변화로 나타남. 즉 설비 변수는 식물 성장에 직접적인 영향을 주지 않음.(광변수는 영향을 주나 데이터의 누락 비율이 높음.)) "펌프상태", "펌프작동남은시간", "최근분무량", "일간누적분무량", "냉방상태", "냉방작동남은시간", "난방상태", "난방작동남은시간", "내부유동팬상태", "내부유동팬작동남은시간", "외부환기팬상태", "외부환기팬작동남은시간", "화이트.LED상태", "화이트.LED작동남은시간", "화이트.LED동작강도", "레드.LED상태", "레드.LED작동남은시간", "레드.LED동작강도", "블루.LED상태", "블루.LED작동남은시간", "블루.LED동작강도", "카메라상태", "냉방온도", "난방온도", "기준온도", "난방부하", "냉방부하", "총추정광량", "백색광추정광량", "적색광추정광량", "청색광추정광량"
* 독립변수의 경우 1분, 종속변수의 경우 1일 단위로 위의 칼럼들의 데이터가 측정되어 제공됨.
* 외부 환경 변수의 경우 설비를 통해 내부 환경변수로 변경되어 청경채에 영향을 미치고 있기 때문에 고려하지 않는다. 추후, 비용을 고려하여 성장 예측을 진행하게 된다면 사용해야 할 수 있으나, 이번 연구에서는 비용적인 부분을 고려하지 않고 환경과 성장에 초점을 맞추고 있어 고려하지 않기로 결정하였다.
* 설비 변수들의 경우 성장에 영향을 끼칠 수 있는 변수들이 led변수들 같이 광량과 관련된 내용과 난방, 분무량, 카메라 상태, 냉/난방 온도 등의 데이터 들이 있음.
* 광량관련 데이터 들의 경우, 다중공선성 등의 문제를 줄이기 위하여 데이터 수가 가장많은 총추정광량 데이터를 사용하였다.
* 설비 변수 들 중에 분무량의 경우 데이터가 적어 제외를 하였고, 온도의 경우 내적 환경 변수와 겹쳐 제거하였으며, 카메라 상태, 난방부하 같은 시설관리 차원의 변수들 또한 제거하였다.
* 최종적으로 "내부온도관측치", "내부습도관측치", "CO2관측치", "EC관측치", "총추정광량"을 선정하여 데이터 분석을 시작하였다.

전처리

* 우선 1일(24시간 \* 60분= 14400분의 데이터를 점을 찍어 그려보았을 때, "EC관측치", "총추정광량" 2개의 변수 변화량이 적어 추후 분석을 진행시에 큰 영향이 없을 것이라고 미리 추정해볼 수 있었다.



* 종속변수와 독립변수의 시간 단위가 달라 독립변수데이터의 시간을 종속변수의 데이터의 단위인 1일로 평균을 통해 맞추었고, 그 전날의 독립변수들에 의해 종속변수의 값들을 예측하는 것이므로 측정일 -1일을 통해 같은 날짜의 데이터로 맞추어 주었다.

**분석결과 정리 보고(analysis result)**

* 독립변수 전체를 포함한 분석

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 회귀분석 통계량 | |  |  |  |  |  |  |  |
| 다중 상관계수 | 0.077429 |  |  |  |  |  |  |  |
| 결정계수 | 0.005995 |  |  |  |  |  |  |  |
| 조정된 결정계수 | 0.003242 |  |  |  |  |  |  |  |
| 표준 오차 | 1.544752 |  |  |  |  |  |  |  |
| 관측수 | 1811 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| 분산 분석 |  |  |  |  |  |  |  |  |
|  | 자유도 | 제곱합 | 제곱 평균 | F 비 | 유의한 F |  |  |  |
| 회귀 | 5 | 25.97821 | 5.195641 | 2.177317 | 0.054167 |  |  |  |
| 잔차 | 1805 | 4307.197 | 2.386259 |  |  |  |  |  |
| 계 | 1810 | 4333.175 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | 계수 | 표준 오차 | t 통계량 | P-값 | 하위 95% | 상위 95% | 하위 95.0% | 상위 95.0% |
| Y 절편 | -0.19681 | 0.362062 | -0.54359 | 0.586793 | -0.90692 | 0.513293 | -0.90692 | 0.513293 |
| 내부온도관측치 | 0.000139 | 0.006942 | 0.020077 | 0.983984 | -0.01348 | 0.013755 | -0.01348 | 0.013755 |
| 내부습도관측치 | -7.8E-05 | 0.000144 | -0.54288 | 0.58728 | -0.00036 | 0.000204 | -0.00036 | 0.000204 |
| CO2관측치 | 0.000333 | 0.000196 | 1.694692 | 0.090306 | -5.2E-05 | 0.000718 | -5.2E-05 | 0.000718 |
| EC관측치 | -2.6E-06 | 0.000128 | -0.0201 | 0.983967 | -0.00025 | 0.000248 | -0.00025 | 0.000248 |
| 총추정광량 | 0.000556 | 0.00021 | 2.644898 | 0.008242 | 0.000144 | 0.000968 | 0.000144 | 0.000968 |
|  |  |  |  |  |  |  |  |  |

전체를 넣고 먼저 분석을 진행해 보았는데 조정된 결정 계수가 0.00324로 매우 낮은 수치를 보이고 있어 이 모델이 유의한지 아닌지 분석하기 전 이미 제대로 된 설명을 하지 못하고 있다는 것을 알 수 있었다.

독립변수별로 p-value를 보았을 때 95%신뢰구간에 속하는 것은 총추정관량 밖에 없고, 90%까지 허용하였을 경우 CO2관측치까지 포함된다. 내부 온도나 습도, EC관측치를 사용하기는 어려워 보인다.

이상치 제거를 통해 CO2 관측치의 경우 어느정도 시도해 볼 만하다고 생각되나 나머지는 포함하기 힘들다고 생각된다.

위와 같은 이유들로 vif를 구하는 의미가 떨어지기는 하나, vif는 다음과 같다.

 VIF 최대값이 1.218486으로 독립변수간 다중공선성은 크게 고려하지 않아도 될 것으로 보인다.

잔차도

텍스트이(가) 표시된 사진

자동 생성된 설명Shapiro-Wilks test를 통해(p-value= 2.2e-6<0.05) 잔차가 정규분포를 따르지 않는다는 것을 알 수 있고, Breusch-Pagan test (heteroscedasticity test)를 통해(p-value= 0.10>0.05) 오차는 이분산이라는 것을 알 수 있었으며, Durbin-Watson Test를 통해 (p-value= 0.83>0.05) 잔차가 자기 상관 관계가 있다고 판단됨. 따라서 잔차들 사이에 자기 상관관계가 있음으로 종속적이라고 판단할 수 있다.

* 텍스트이(가) 표시된 사진

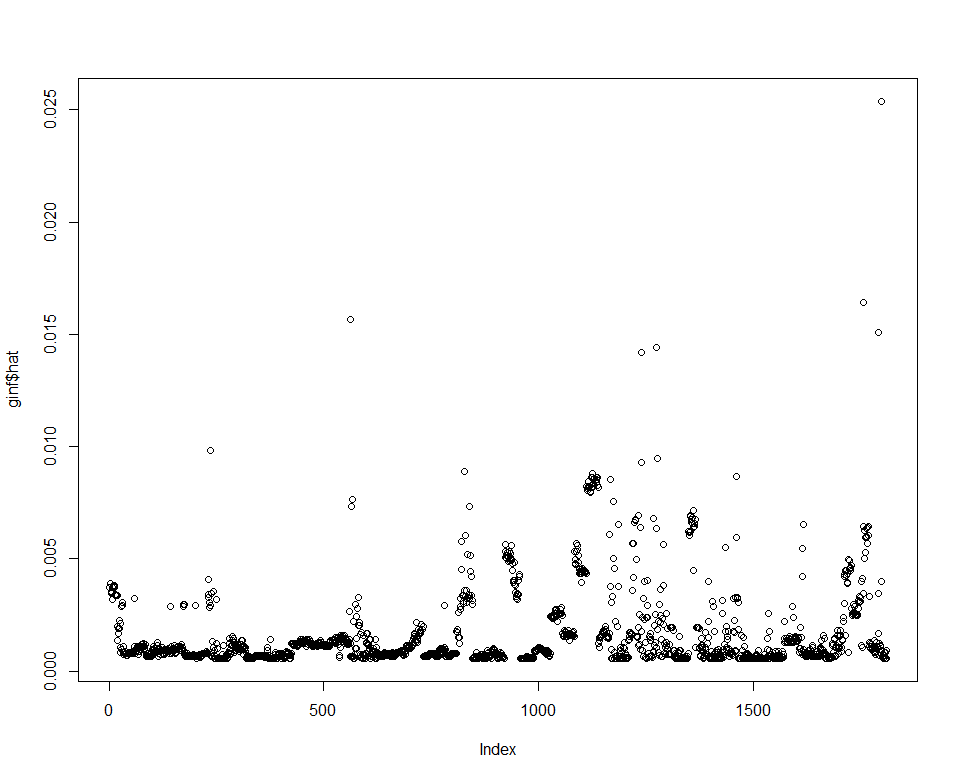
  자동 생성된 설명독립변수 CO2 관측치, 총추정광량과 종속변수 간의 분석

유의 수준 95%로 판단하였을 경우 총추정관량 만이 유의하며 유의 수준을 90%로 잡을 경우 CO2관측치까지 유의해진다. 다만 조정 결정 계수가 0.004732로 여전히 매우 낮은 수치를 보이고 있어, 모델이 제대로 청경체의 성장을 설명하지 못한다고 판단됨에 따라, 이 분석의 의미가 있는가에 대한 의문을 갖게 된다.

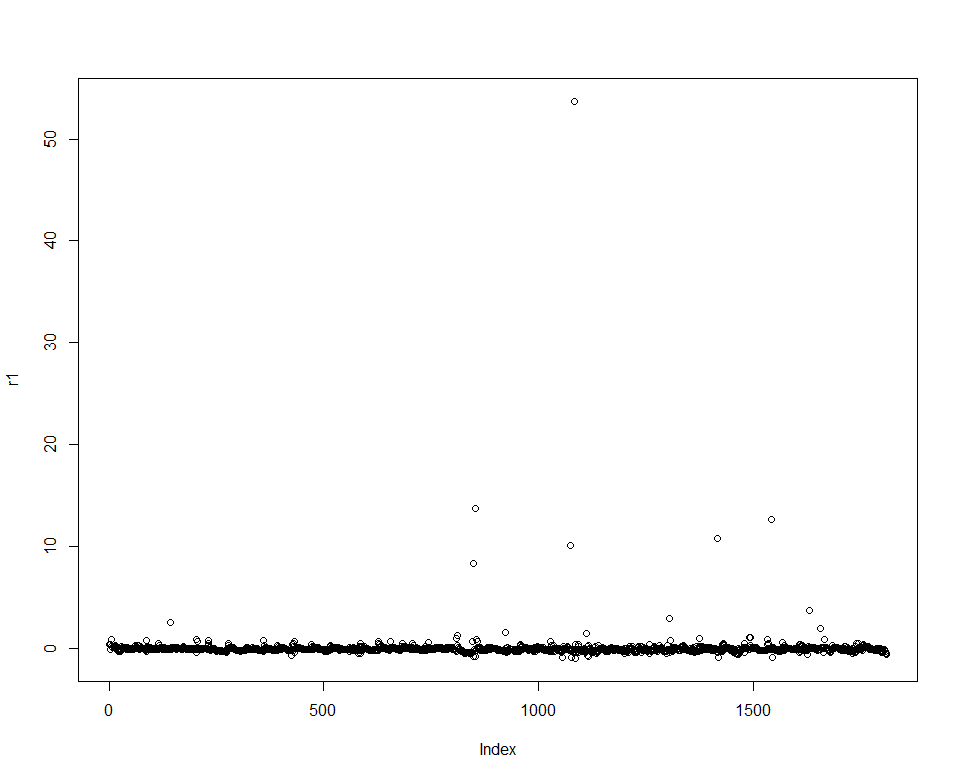
텍스트이(가) 표시된 사진

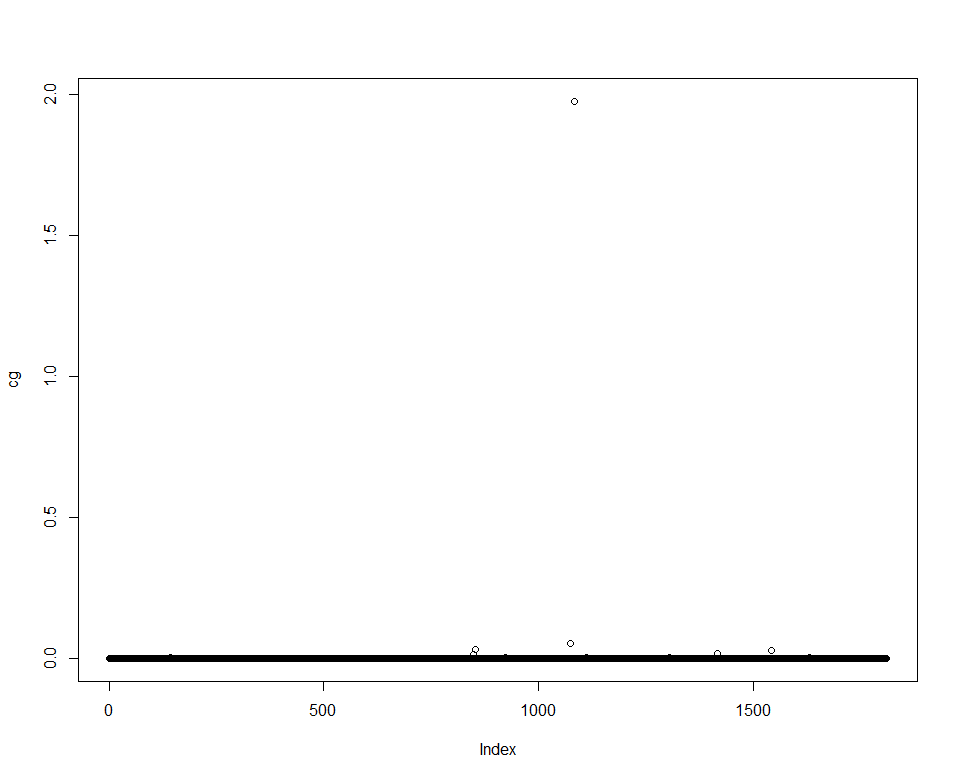
자동 생성된 설명이전의 분석을 통해 상관성이 낮다는 것을 알고 있지만 다시 진행해보았을 때, 이전보다 더욱 낮은 수치들이 나왔다. 물론 그 전에도 1에 가까운 수치들이었기 때문에 의미있는 변화는 아니라고 판단된다.

Shapiro-Wilks test를 통해(p-value= 2.2e-6<0.05) 잔차가 정규분포를 따르지 않는다는 것을 알 수 있음.

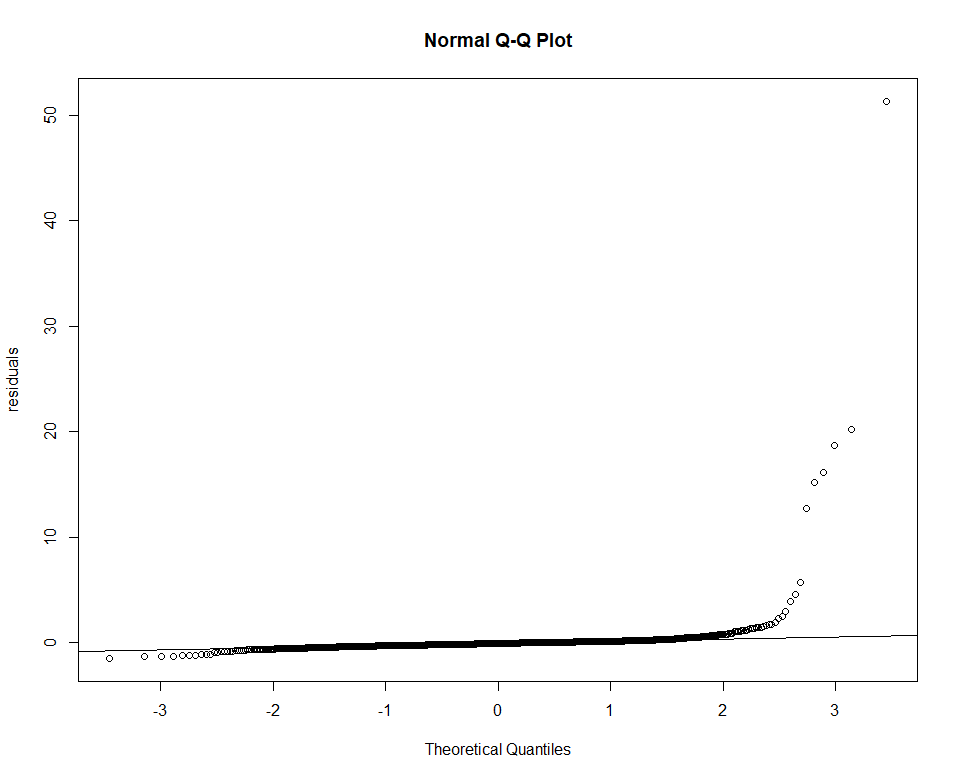
텍스트이(가) 표시된 사진

자동 생성된 설명Breusch-Pagan test (heteroscedasticity test)를 통해(p-value= 0.012<0.05) 오차는 동분산이라는 것을 알 수 있었으며, Durbin-Watson Test를 통해 (p-value= 0.84>0.05) 잔차가 자기 상관 관계가 있다고 판단됨. 따라서 잔차들 사이에 자기 상관관계가 있음으로 종속적이라고 판단할 수 있다.

Leverage point를 이용하여 이상치가 있는지 확인하였을 때 평균의 3배이상인 값은 없는 것으로 확인되었음.

studentized deleted residuals을 이용하여 이상치가 있는지 확인하였으나, 발견되지 않았음.

영향치(influemtial point)가 있는지 확인하였을 때 

1084쨰 값이 확인되었음

좌측에서 산포도 점들이 q-q라인 보다 아래쪽에 있고, 우측에서는 위로 있어, 좌우로 long-tail 형태의 분포를 가진다는 것을 알 수 있었음.

결론적으로, 90% 신뢰구간에서

‘청경채 잎 면적 증감률 = 0.0005483 \*총추정광량 + 0.0003341\* CO2관측치 -0.1993381’

임을 알 수 있었으나 설명력 부분이 너무 낮아서 생육환경을 조성하는데 참고 모델로 사용하기에는 무리가 있다고 판단되고, 유전적 요소, 공기 순환, 비료 등 다양한 요소들이 복합적으로 영향을 끼쳐 성장이 결정되는 것이지 단순하게 몇 가지 통제할 수 있는 변수로 작물을 키울 수 있는 것이 아닌 것 같다는 생각을 하게 되었다. 기존 논문들을 살펴볼 때도 유의미한 결과를 낸 것이 없어서 데이터 분석을 진행할 때 불안요소로 남았는데,

**참고문헌(reference list)**

- “스마트팜 데이터를 이용한 토마토 최적인자에 관한 연구”, 한국데이터정보과학회지. 2017-11, 나명환, 박유하, 조완현

- “머신러닝 기반 식용 작물별 생산량 예측 연구”, 성균관대학교 일반대학원, 2020, 박성호

* "Crop Yield Prediction Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, A. Nigam, S. Garg, A. Agrawal and P. Agrawal,:

**데이터 출처(data source)**

KIST 강릉분원 천연물연구소-천연물라이브러리(https://gn.kist.re.kr/portal/main/main.do)