**예측애널리스틱 과제 1**

**2020170850 김형진**

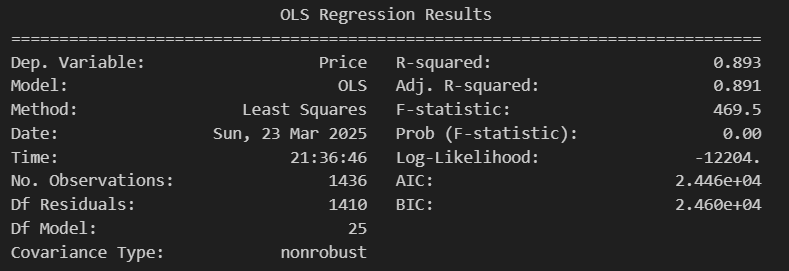
**과제내용**

**첨부한 ToyotaCorolla 데이터를 이용하여 다음 분석을 실시하시오.**

* **다중회귀모델 구축 및 해석 (R2, Adjusted R2  포함)**
* **각 회귀계수에 대한 95% 신뢰구간 추정**
* **각 회귀계수에 대한 기울기=0 여부 검정**
* **데이터를 Training set과 Testing set 으로 나누고 Testing set의 예측정확도 계산**
* **ToyotaCorolla 데이터 분석 관련 스스로 1문제 내고 푸시오.**
* **고려대학교 교훈, 상징동물, 교목, 교색깔, 교가(1절만)를 적으시오.**

**다중회귀모델 구축 및 해석 (R2, Adjusted R2 포함)**

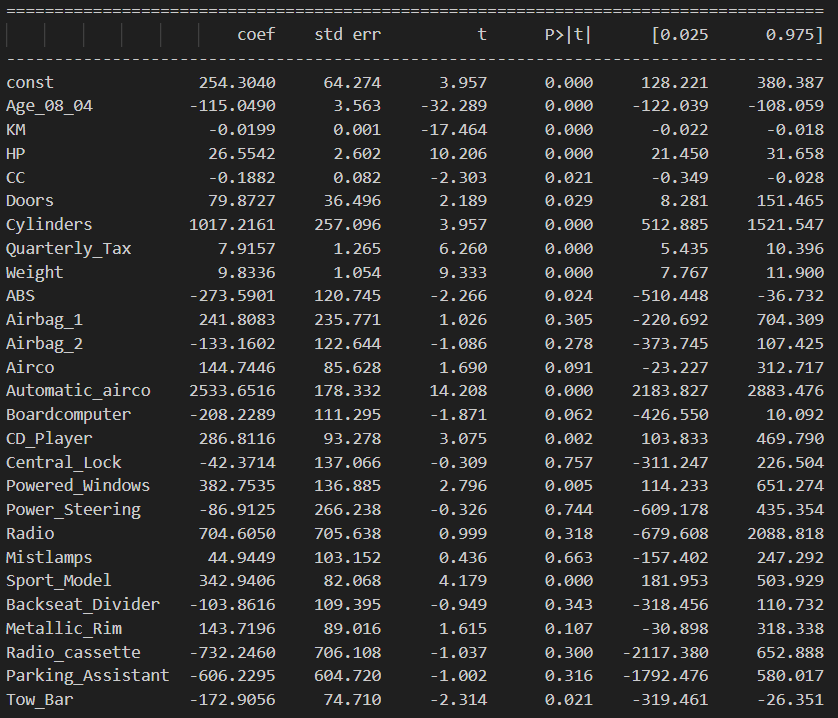
ToyotaCorolla.csv 데이터에서 ‘Price’를 종속변수(y)로 지정하고 나머지 변수들(‘Id’, ’Model’은 index columns이라 제외)을 독립변수(X)로 지정하여 다중회귀모델을 구축.



단순히 y(‘Price’)의 평균으로 y를 예측했을 때보다 이 모델을 사용하여 X를 통해 y를 예측했을 때 설명력이 89.3% 좋아졌다.

R2, Adjusted R2값을 비교했을 때 0.002 차이밖에 안 나는 것을 보아 현재 사용한 변수들이 대체로 유의미하며, 불필요한 변수가 많지 않다고 볼 수 있다.

**각 회귀계수에 대한 95% 신뢰구간 추정**



**각 회귀계수에 대한 기울기=0 여부 검정**

각 회귀계수의 점 추정치가 0이라는 귀무가설을 세운 뒤 이를 통해 t값을 구하고 P-value를 측정한 값을 보니 유의수준 0.05보다 큰 변수는 다음과 같다.



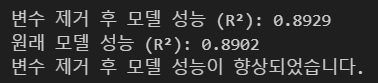
이 변수들은 회귀계수의 기울기가 0이라는 귀무가설을 유의수준 0.05 수준에서 기각할 수 없다.

**데이터를 Training set과 Testing set 으로 나누고 Testing set의 예측정확도 계산**

데이터를 80% Train / 20% Test로 분할하여 Train data로 회귀모델을 학습한 후 Test data에서 예측값을 구했다. Test data에서 예측값과 실제값을 비교하여 예측 정확도(R²)를 측정하였더니 0.890이 나왔다.

**ToyotaCorolla 데이터 분석 관련 스스로 1문제 내고 푸시오.**

새로운 문제: 만약 회귀계수의 기울기가 0이라는 귀무가설을 기각하지 못한 변수를 제거하고 다중회귀모델을 학습하여 예측했을 시 성능을 구하시오.



P-value가 0.05 이상인 변수들 제거하여 모델을 훈련하고 평가하였더니 원래 모델보다 성능이 소폭 상승했다. P-value가 큰 변수를 제거함으로써 모델의 복잡도가 줄어들고 과적합이 감소하면서 성능이 향상되었다고 볼 수 있다.

**고려대학교**

**교훈: 자유, 정의, 진리 (LIBERTAS, JUSTITIA, VERITAS)**

**상징동물: 호랑이**

**교목: 잣나무**

**교색깔: 크림슨**

**교가:**

**북악산 기슭에 우뚝 솟은 집을 보라**

**안암의 언덕에 퍼져나는 빛을 보라**

**겨레의 보람이요 정성이 뭉쳐**

**드높이 쌓아 올린 공든 탑**

**자유 정의 진리의 전당이 있다**