<Quiz 12>

이번 시간에는 Boston house price 데이터(이하 Boston 데이터)를 분석해보고 선형회귀 모델을 이용해 집값 예측을 해보고자 한다.

1. Boston 데이터를 분석하기에 앞서 기초적인 회귀 모델을 만들어보자.

1) Random Number Generator 0번을 이용하여 [0, 2] 구간에 대응되는 uniform distribution으로부터 샘플링 한 100 x 1 크기의 변수 x를 생성하고, 다음과 같은 관계를 가지는 변수 y를 생성하도록 하자.

여기서 은 평균이 0이고 표준편차가 1인 정규분포로부터 샘플링 한 랜덤 벡터이다.

2) 앞서 생성한 x와 y에 대해 다음과 같이 좌표 평면 상에 (x, y)를 도시해보자.



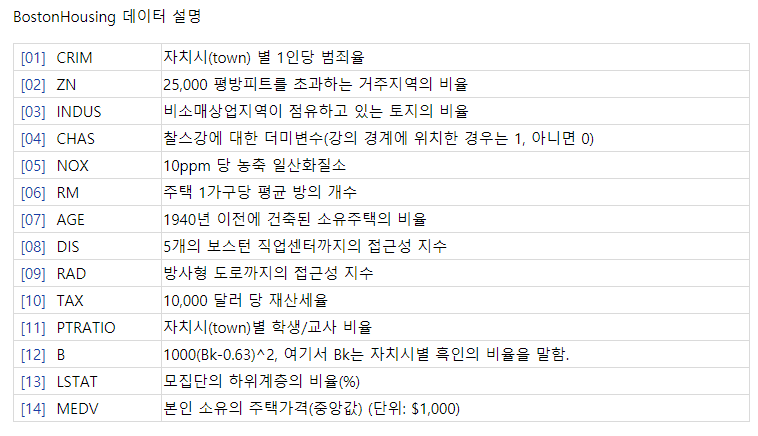
3) 위의 (x, y) 데이터셋에 대해 polyfit 함수를 이용해 1차 다항 함수 선형회귀 모델을 결정해보고 선형회귀모델의 수식을 확인해보자.

4) 앞서 3)에서 결정한 선형회귀 모델을 아래와 같이 도시해보자.



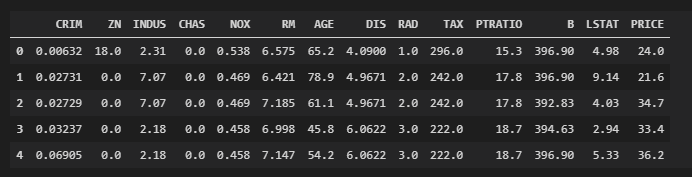
2. Boston 데이터를 불러오고 회귀모델을 만들어 보자.

Boston 데이터는 총 13개의 feature를 가지고 있으며, 각 feature들의 의미는 아래와 같다.



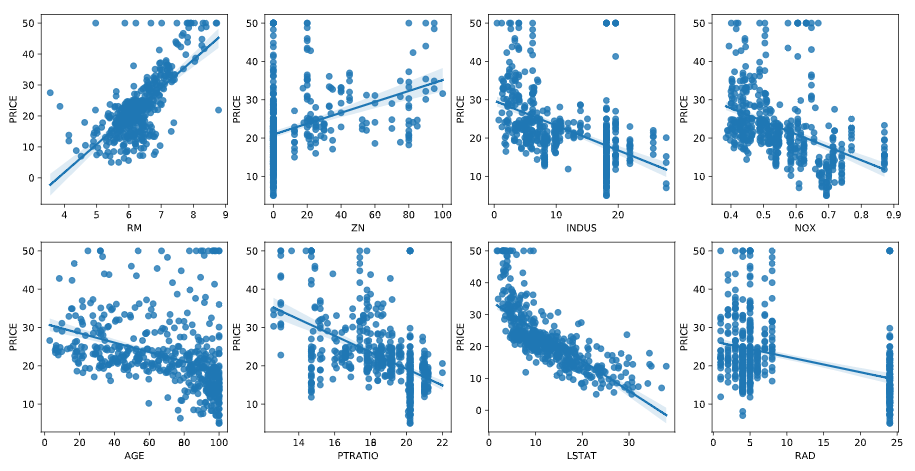
(<http://dator.co.kr/?vid=ctg258&mid=textyle&document_srl=1721307>)

우리의 Boston 데이터는 ‘boston.mat’ 파일에 들어있으며 최초 5개 row의 데이터들은 아래와 같다.



1) 데이터를 불러오고 위의 그림과 같은 결과를 확인할 수 있는지 점검해보도록 하자

2) Boston 데이터를 이용해 할 수 있는 일의 핵심은 어떤 feature가 집값을 예측하는데 좋은 factor로 작동하는지를 알아내는 것이다. 우리는 RM, ZN, INDUS, NOX, AGE, PTRATIO, LSTAT, RAD 의 총 8개 feature와 집 값 간의 관계를 확인하기 위해 아래와 같이 8개의 subplot 위에 각 feature와 집 값(price)에 대한 산점도(scatter plot)을 그리고 선형 회귀 모델을 구하여 함께 도시해보자.



3) 이 결과를 보면 방 개수(RM)과 LSTAT(하위 계층의 비율)이 중요한 factor로 작용한다는 것을 알 수 있다. RM은 양 방향의 선형성이 가장 큰 feature이고, LSTAT은 음 방향의 선형성이 가장 큰 feature임을 알 수 있다. 이 말이 의미하는 것은 방의 개수가 많을수록 집 값이 올라가고, 하위 계층의 비율이 적은 곳에 있는 집일수록 집 값이 올라간다는 것을 말한다.

각 feature들의 선형성의 정도를 확인하기 위해 각 feature에 대한 선형회귀 모델의 slope 값을 모두 확인하자.

3) 이번에는 Boston 데이터 셋에 있는 13개의 모든 feature를 한꺼번에 이용하여 다변수회귀모델을 만들어보자. MATLAB에서 다변수 회귀모델을 만들어주는 함수는 mvregress()이다.

