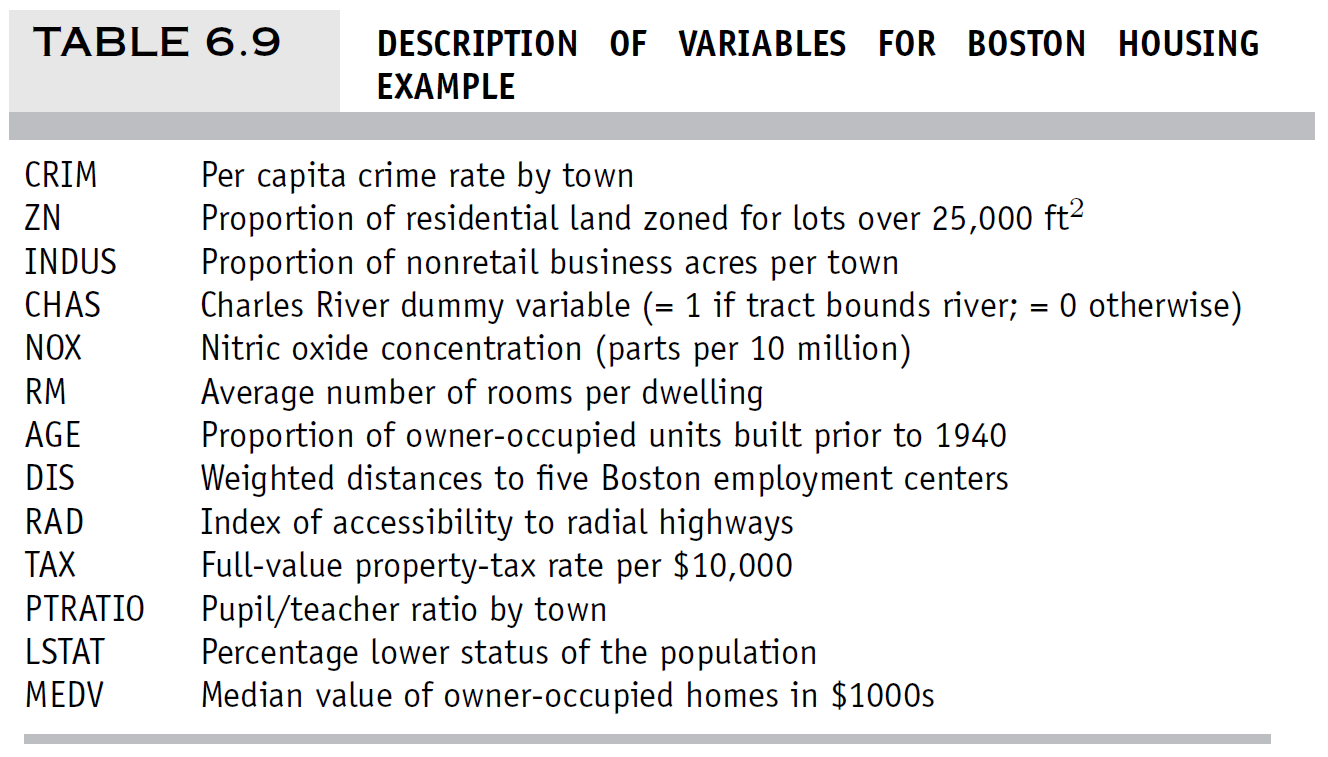
**[일곱 번째 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

**6.1 보스턴 주택가격 예측 BostonHousing.csv 파일은 미국 인구통계조사국에 의해 수집된 매사추세츠주 보스턴 지역의 주택정보를 담고 있다. 이 데이터세트는 보스턴 지역의 506개의 인구통계조사 주택단지 정보를 포함한다. 데이터 수집의 목표는 범죄율, 공해, 방의 개수에 대한 정보에 근거하여 새로운 주택단지의 주택가격의 중앙값(median)을 예측하는 것이다. 데이터세트는 13개의 예측변수와 반응(목표)변수인 주택가격의 중앙값(MEDV)을 포함한다. <표6.9>는 예측변수들과 반응변수를 설명하고 있다.**



**a. 데이터를 학습 세트와 검증 세트로 나누는 이유는 무엇인가? 학습 세트와 검증 세트는 어떠한 용도로 사용되는가? CRIM, CHAS 및 RM를 사용하는 함수로서 주택가격 중앙값에 대한 다중선형회귀모델을 만드시오.**

1. 데이터를 학습 세트와 검증 세트로 나누는 이유는 무엇인가?

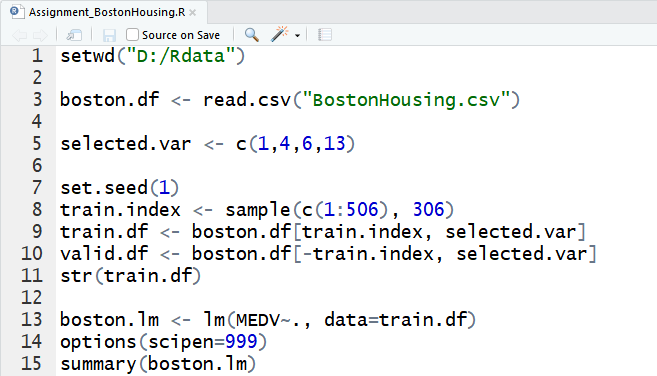
* 새로운 데이터에 모델이 얼마나 잘 일반화되는지 측정하기 위해서이다. 즉 모델이 학습 세트에 잘 맞는 것보다, 학습 과정에 없던 새로운 데이터에 대해 얼마나 예측을 잘하는지 확인하기 위해서이다. 또한 과적합을 방지하기 위해서이다.

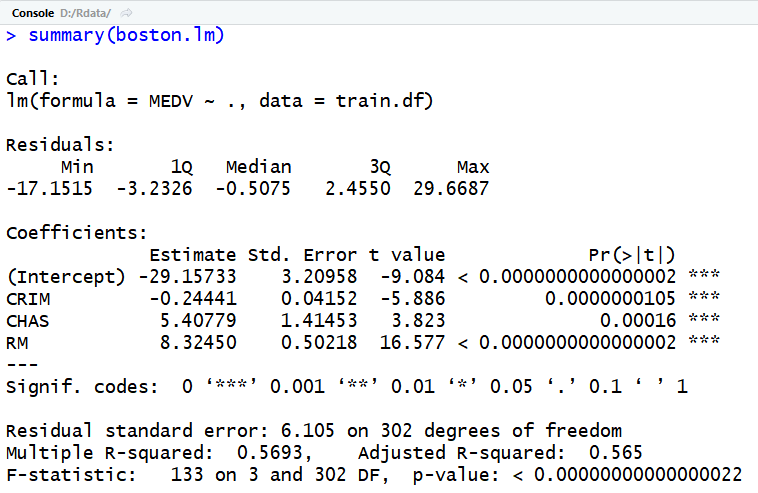
1. 학습 세트와 검증 세트는 어떠한 용도로 사용되는가?

* 학습 세트는 모형을 적합화시키기 위한 용도로 사용된다.

검증 세트는 모형이 얼마나 잘 적합화 되었는지 평가하고, 일부 모형들을 조정하며, 구축된 모형들 중에서 가장 좋은 모형을 선택하기 위한 용도로 사용된다.

1. CRIM, CHAS 및 RM를 사용하는 함수로서 주택가격 중앙값에 대한 다중선형회귀모델을 만드시오.





<다중선형회귀모델>

**b. 모델의 예측변수들로부터 주택가격 중앙값을 예측하기 위한 식을 쓰시오.**

* 위 문제a. 결과를 이용하여 주택가격 중앙값을 예측하는 식은 다음과 같다.

**c. 찰스강 경계에 위치하지 않고, 범죄율이 0.1이고, 평균 방의 개수가 6개인 보스턴 주택단지의 주택가격 중앙값은 얼마로 예측되는가? 예측오차는 얼마인가?**

* 위 문제b. 결과를 이용하여 CRIM = 0.1, CHAS = 0, RM = 6 대입하면 값은 다음과 같다.
* 따라서 찰스강 경계에 위치하지 않고, 범죄율이 0.1이고, 평균 방의 개수가 6개인 보스턴 주택단지의 주택가격 중앙값은 약 20.76523으로 예측된다.
* ~~위 문제b. 다중선형회귀모델을 통해 예측오차는 6.105 인 것을 알 수 있다.~~

**실제로 예측오차는 구할 수 없다. 6.105는 표준오차이다.**