DS 회귀분석 팀과제

* 공통 부분 외에 선택 1 또는 선택 2 중 택일, 과제 앞부분에 선택을 명기할 것

1. Reference
   1. <https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html>
   2. <https://daviddalpiaz.github.io/appliedstats/variable-selection-and-model-building.html>
2. 데이터 자유선택 단, 아래 데이터 제외   
   https://archive.ics.uci.edu/ml/datasets/Housing. (Boston Housing) - 대표적인 회귀용 데이터, 연습용  
   https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees (봉제 노동자 효율성)  
   https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand (서울 자건거 대여)
   1. 연습에 도움이 되는 적절한 크기, 복잡성을 가지는 데이터 선택
   2. 데이터에 대하여 설명, 수집 목적, 변수 설명
3. 위 데이터 셋을 기반으로 목표변수을 결정하는 선형회귀 모델을 생성하라.
   1. 예측, 설명의 목적으로 각각 모델을 생성한다.
   2. 각 변수의 회귀계수의 부호와 양에 대하여 해석
4. 예측모델로써의 성능을 측정하라.
   1. 테스트 집합에 대한 RMSE 사용 – CV 필요
5. 설명모델로써의 성능을 측정하라.
   1. 테스트 집합이 필요 없음. – 이유는?
   2. 설명모델로써의 성능과 유의미한지를 test 하라.
      * 모델 F-test 의 p-value, R, Adjusted-R 등을 이용한다.
   3. 목표변수에 영향을 미치는 강도에 따라 변수들을 순서대로 나열해 보라.
      * 변수에 대한 T-test 의 p-value를 기준으로 한다
   4. (선택 1) 변수 선택 Stepwise-selection을 실행
      * forward / backward 방법에서 각각 먼저 선택 / 나중에 제거 되는 변수 일수록 더 중요한 변수라고 할 수 있다.
      * 위 항목의 p-value 와 비교해 볼 것
   5. 각 변수의 회귀계수는 어떠한 의미를 가지는가?
      * 설명력과는 무관함.
      * 단지 비례관계(증감의 정도 및 방향)의 상수를 나타냄.
      * 영향력이란 예측 변수와 목표변수와의 연관도를 나타낸다. 🡪 p value
6. (선택 2) 다른 회귀 모델과 비교
   1. 결정나무, random forest, 신경망 등을 사용하여 회귀 모델 생성
   2. 정확도, 변수 중요도 등의 측면에서 비교

- 끝 -