(공유)회귀분석 (본교과목명: 자료분석 및 실습)

자료분석과정에 대한 이해와 더불어 배운 방법론의 실질적 문제해결을 위한 적용을 경험한다. 과제를 통해서 문제를 설정하고, 실제 데이터 분석을 통해 답안을 도출하는 과정을 배운다. 이를 통해 학생들은 이론으로 배운 내용들을 변형하고 응용하여 알맞은 모델을 제작할 수 있는 문제해결능력을 기르게 된다. 이론과 실습의 복합을 통해 통계적 방법론들의 강점과 한계를 쉽게 이해하고 비판적 사고 능력을 향상시킨다.

Course Information

Instructor: 이권상 (자연과학대학 통계학과)

Class time: 화, 목 11:00-12:50

Office hour: by appointment

TA: TBD

Course Objectives

해결하고자 하는 문제의 구체화를 포함하여, 데이터의 수집 및 정리 (data collection and cleaning), 탐색적 데이터분석 (exploratory data analysis), 시각화 (visualization), 통계적 추론 및 예측 (statistical inference and prediction), 그리고 의사결정 (decision-making)의 핵심원리를 배운다. 또한 실습문제를 통해 일련의 과정을 경험한다.

자료분석 및 실습은 크게 두 가지 주제로 이루어져 있다.

1. 데이터의 이해/시각화: 문제해결을 위한 데이터의 수집 방법에 대한 이해를 높이며, 포함된 변수들이 사용하고자 하는 통계모형에 적합한지를 판단하는데 도움을 준다. 또한 시각화방법을 통해 사용된 모형의 진단도 가능하게 한다.

- 탐색적 데이터 분석
- 데이터의 정리 및 요약
- 통계모형의 진단

2. 통계적 방법론의 이해: 목적에 맞는 통계적 방법론을 공부하고, 사용되는 모델들의 기본 원리를 배운다.

- 지도학습(supervised learning): 회귀분석모형(regression), 분류모형(classification), 의사결정트리모형(decision tree)

- 비지도학습(unsupervised learning): 주성분 분석(principal component analysis, PCA), 군집화모형(clustering)

Prerequisite

선수과목은 정해져 있지 않지만, 기본 통계학의 개념과 통계적방법론에 대한 이해가 요구됩니다. 자료분석을 위해 R 프로그래밍의 기초도 요구됩니다.

Course Materials

1. Modern Data Science with R, 2nd edition by Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton. (https://mdsr-book.github.io/mdsr2e/)
2. Practical Statistics for Data Scientist, 2nd edition by Peter Bruce, Andrew Bruce, and Peter Gedeck (또는 한글판: 데이터 과학을 위한 통계 2 판)

Grading/Evaluation

기말프로젝트 50%, 과제 40%, 그리고 수업태도/참여 10%로 평가한다. 기말프로젝트는 마지막 두 주에 걸친 프로젝트발표와 리포트 점수를 합해서 결정된다. 기말프로젝트를 위한 기본적인 내용의 습득을 위해 과제도 함께 주어진다.

Course Plans

Week 1	Introduction (overview, sampling and probability)
Week 2	Data wrangling
Week 3	Exploratory data analysis (EDA) and Regular expression (regex)
Week 4	Visualization I, II

Week 5	Data science ethics and Modeling
Week 6	Linear regression (simple linear regression and OLS)
Week 7	Feature engineering and Bias & Variance
Week 8	Cross-Validation (CV), Regularization, and Gradient descent.
Week 9	Logistic regression I, II
Week 10	Classification I, II (regularization, multiclass classification)
Week 11	Decision trees I, II
Week 12	Inference for modeling and PCA
Week 13	Clustering I, II
Week 14	Final project presentation
Week 15	Final project presentation