Inference for statistics and data science course

Take-home assignment 2024/2025

Choose an assignment between Option A and Option B and submit the report before Sunday, January 5th 2025 23:59.

Option A - Prediction modelling

The aim of this assignment is to develop a prediction model for 30-day mortality after an acute myocardial infarction, from a data set: containing 785 patients, 52 of whom died.

The dataset is available here.

Overview of the data set

A total of 17 predictors is considered.

Variable	**Explanation**
Day30_mortality	30-day mortality (0/1)
Gender	Female gender (0/1);
Age	Age in years (range: 19-110)
Killip_class	Killip class: a measure for left ventricular function; class 1-4
Diabetes	Diabetes (0/1)
Hypotension	Hypotension (systolic BP<100) (0/1)
Heart_rate	Heart rate (tachycardia: pulse>80) (0/1)
Anterior_infarct_location	Anterior infarct location (0/1)
Previous_myocardial_infarction	Previous myocardial infarction (0/1)
Height	Height in cm (range: 140-212)
Weight	Weight in kg (range: 36-213)
Hypertension	Hypertension history (0/1)
Smoking	Smoking (1=never; 2=ex; 3=current smoker)
Hypercholesterolaemia	Lipids: hypercholesterolaemia (0/1)

Previous_angina_pectoris	Previous angina pectoris (0/1)
Family_history_of_MI	Family history of MI (0/1)
ST_elevation_leads	ST elevation on ECG: number of leads (range: 0-11)
Time_To_Relief	Time to relief of chest pain > 1 hour (0/1)

Assignment

Develop a model from the provided data using R or Python and write a report explaining the model development process. Use different techniques to derive the model in different ways so you can pick the best one.

Use the TRIPOD checklist (available here) to guide your reporting (elements 8-20). Make sure you report:

- Any preprocessing done to the data (e.g. cleaning, class imbalance)
- How you handled missing data
- Any statistical test ran on the data
- Which statistical or machine learning techniques you used to derive the model
- Feature selection (if any)
- Hyperparameter tuning (if any)
- Estimates of model performance (and techniques used to estimate it)
- Interpret the results of the model (e.g. try to explain comparative results).

Make sure you describe and justify the decisions you made during model development and to provide estimates of your model's performance using internal validation. Include the code you used for the analyses as an appendix.

Deliverable

Report including source code.

Option B - Causal inference

The aim of this assignment is to derive different estimates of the effect of an intervention. We will use a dataset from an observational study that aimed to evaluate a "nudge-like" intervention to change student behavior. The main goal of the study was to assess the heterogeneity in the effect of the intervention.

The data is available here. Please note the data has already been preprocessed.

Variable	**Description**
Outcome	Measure of student's performance after intervention (continuous variable)
Intervention	Whether the student received the intervention (0= no, 1=yes)
expectations_future	Student's expectations for success in the future
ethnicity	Student's race/ethnicity (Categorical variable)
gender	Gender (Categorical variable, 1=male, 2=female)
first_generation	Student's first-generation status (i.e. first in family to go to college) (Categorical variable)
school_urbanity	Urbanity of school (i.e. rural, suburban, etc.)
school_fixed	School-level % of students that believe that intelligence is a fixed trait
school_achievement	Past school achievement level
school_ethnic	School racial/ethnic minority composition (% black, latino, or native/american)
school_poverty	School poverty concentration (% of students below the poverty line)
school_size	School size

Assignment

Analyse the data and write a report answering to the following questions:

1. Was the intervention effective in improving student performance? If so, how effective was it?
2. How did the prior school achievement affect the effectiveness of the intervention?
3. Were there any other variables that affected the effectiveness of the intervention? If so, how?

Please justify your responses to the above questions describing the analyses.Try to use a range of different estimators (and techniques) that we learned in class. You can also try to implement more advanced estimators. Make sure you describe any preprocessing or statistical test you run on the data.

Deliverable

Report including source code. Include the code you used for the analyses as an appendix.