Practicals Day 1 Afternoon Session

PhD course: Causal prediction for medical decision making

Exercise 2

We have synthesized data alike the training data behind the Steno type 1 risk engine. The data are not real data but obtained by computer simulation for the sole purpose of illustrating the methods discussed during the workshop. Since we have simulated the data we have removed the obstacles from censoring and provide a binary outcome variable which indicates if CVD occured before the 5 years horizon where no occurrence means either death without CVD or alive without CVD.

Study population Adults with type 1 diabetes. It is recommended for patients to come for regular follow-up visits every 4 months. These data represent information available from the past year at the time that the predictions are to be made (time 0), which includes some measurements at 2 time points. The variable pid is a unique patient identifier.

Time zero Date of a screening visit at diabetes center

Outcome Cardiovascular disease within 5 years of time 0. The is coded as 0 for no, 1 for yes, in the variable $\text{cvd}_{\text{5year}}$

Predictors all measured at or before time 0

- `age`: Age in years at the time 0
- `sex_male`: Binary indicator of male sex (0 = female, 1 = male)
- `diabetes_duration`: Time in years since diagnosis of type 1 diabetes
- `smoking`: Binary indicator of current smoking status (0 = non-smoker, 1 = smoker)
- `motion`: Binary indicator of regular physical activity (0 = sedentary, 1 = some moderate physical activity)

- `HBA1C_time1`, `HBA1C_time2`: Glycated hemoglobin, also called hemoglobin A1C value at two time points in mmol/mol. Smaller values indicate better glycemic control.
- `urine_albumin_time1`, `urine_albumin_time2`: The urine albumin to creatinine ratio, in mg/g. A larger relative concentration of albumin in the urine is an indicator of poor kidney function
- `LDL_time1`, `LDL_time2`: Low-density lipoprotein measurement in serum, in mmol/L.
- `SBP_time1`, `SBP_time2`: Systolic blood pressure in mmHg
- `eGFR_time1`, `eGFR_time2`: Estimated glomerular filtration rate, in $mL/min/1.73m^2$. An estimate of kidney function, larger values means better kidney function
- `statin`: indicator of statin use between times 1 and 2 (0 = not on statins, 1 = on statins)

Perform the following steps in R:

0. Load R-package `riskRegression`
1. Read the train and the test data into R
2. What is the prevalence of CVD 5 years after time zero?
3. Note that the null model ignores the covariates and predicts the same risk to everyone. It can be obtained with the `glm` function as follows:
4. Fit a bench mark model using logistic regression including additive effects of sex, age and diabetes duration
5. Fit a model which resembles the formula of the Steno 1 risk engine using logistic regression
6. Does the Steno model outperform the benchmark model in the train data? Use 10-fold cross-validation
7. Does the Steno model outperform the benchmark model in the test data?
8. Use any method to make a new risk prediction model and compare it to the Steno model on the test data. Note that the Score function allows you to specify a vector with predicted risks. Here is an example using random forests based on the R-package ranger.