Project - Applied Statistical Learning - 2024

Karin Groothuis-Oudshoorn & Julia Mikhal

Project Description: Prediction of Secondary Cardiovascular Events

The *data* for this project is taken from a prospective cohort study, which took place from September 1996 to March 2006. In total 3873 patients, who had a clinical manifestation of atherosclerosis were enrolled in this study.

The *study* was designed to:

- establish the prevalence of concomitant arterial diseases and risk factors for cardiovascular disease in a high-risk population
- identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease.

The *aim* of this project is to develop prediction models for the long-term outcome of a cardiovascular event.

The following variables (in total 31) are in the data set:

outcomes:

- `EVENT`: Presence of cardiovascular event (*1 = Yes, 0 = No*)
- `TEVENT`: Number of *days* the patient is in the study till the event occurs or the patient is censored

patient descriptives:

- `SEX`: Gender (*1 = male, 2 = female*)
- `AGE`: Age, *years*
- `LENGTH`: Length, *m*
- `WEIGHT`: Weight, *kg*

classical risk factors:

- `BMI`: BMI, *kg/m2*
- `DIABETES`: Diabetes (*1 = Yes, 0 = No*)
- `SMOKING`: Smoking (*1 = never, 2 = former, 3 = current*)
- `PACKYRS`: Number of pack years
- `ALCOHOL`: Alcohol (*1 = never, 2 = former, 3 = current*)
- `SYSTBP`: Blood pressure, systolic automatic, *mm Hg*
- `DIASTBP`: Blood pressure, diastolic automatic, *mm Hg*
- `SYSTH`: Blood pressure, systolic by hand, *mm Hg*
- `DIASTH`: Blood pressure, systolic by hand, *mm Hg*
- `CHOL`: Total cholesterol, *mmol/L*
- `HDL`: High-density lipoprotein cholesterol, *mmol/L*
- `LDL`: Low-density lipoprotein cholesterol, *mmol/L*
- `TRIG`: Triglycerides, *mmol/L*

previous symptomatic atherosclerosis:

- `CEREBRAL`: Previous symptomatic atherosclerosis, cerebral artery disease (*1=Yes, 0=No*)
- `CARDIAC`: Previous symptomatic atherosclerosis, coronary artery disease (*1=Yes, 0=No*)
- `AAA`: Previous symptomatic atherosclerosis, abdominal aortic aneurysm (*1=Yes, 0=No*)
- `PERIPH`: Previous symptomatic atherosclerosis, peripheral arterial disease (*1=Yes, 0=No*)
- `HISTCARD`: Sum score of previous symptomatic atherosclerosis (`CEREBRAL + CARDIAC + AAA + PERIPH`)
- `HISTCAR2`: Sum score of previous symptomatic atherosclerosis (`CEREBRAL + CARDIAC + 2*AAA + PERIPH`)

markers of atherosclerosis:

- `HOMOC`: Markers of atherosclerosis, homo cysteine, μ *mol/L*
- `GLUT`: Markers of atherosclerosis, glutamine, μ *mol/L*
- `CREAT` Markers of atherosclerosis, creatinine clearance, *mL/min*
- `IMT`: Intima media thickness, *mm*
- `ALBUMIN`: Albumin (*1 = no, 2 = micro, 3 = macro*)
- `STENOSIS` : Carotid artery stenosis $> 50\%$ (*1 = Yes, 0 = No*)

There are *two associated endpoints*: `EVENT` and `TEVENT` in this data set. The most straightforward model you can build is a classification model based on `EVENT` only. Alternatively, you can develop a survival model based on both `EVENT` and `TEVENT`. `TEVENT` contains the time (in *days*) when the event occurred or how long a patient was observed in the cohort and censored at the end (so, no event occurred). Additionally, *two data sets* are provided: one without missing data (**SMARTc.csv**) and one with missing data (**SMART.csv**).

Project Assignment

General Information:

- For the whole project you work in a group of 2-3 students.
- The project assignment consists of 4 parts, which are described below. For each part, the deadline for submission is specified.
- The data set with missing values (SMART.csv) is used only once (for question (d) in Part 1). For all other parts of the assignment, please use the data set without missing values (SMARTc.csv). Read the questions carefully and give complete answers.

Deliverables:

- For each part of the assignment, upload 2 files: the Rmarkdown (or Quarto) file and the PDF (or HTML) file in which the R-code, results, figures, and text are visible. For the reproducibility of your calculations, please, specify your random seed, and include all libraries you used.
- The files should be named as follows: groupX_partY.Rmd (or .qmd) and groupX_partY.pdf (or .html). Here X is your group number and Y is the number of the part of the project.

Grading:

- Grading: the maximum number of points for the project is 100. The grade for the project is the number of points divided by 10.
- For every late submission for more than one of the deadlines, 10 points will be subtracted.
- If you have less than 50 points for your project you can resubmit it with your group. The deadline is 2 weeks after the end of the quartile. Your group will be then invited to discuss the project

Remarks:

- If you have to compare outcomes, take into account the uncertainties of these outcomes (e.g., confidence intervals).
- For assessing the performance of a model consider the methods from Chapter 5 and the extra notes on validation.
- You don't need to write a whole report with an introduction and discussion. Write it in a question/answer format.

PART 1: (deadline: September 16, 17:00) [25 points]

In this part, you will describe the data set, visualize the variables, estimate a logistic regression model for the entire data set, and apply the imputation of the missing values.

(a) Provide a short written description of the data set including a table of the variables from the data set (outcome variables and predictors) using suitable descriptive statistics [**5 points**].

(b) Create **two different** types of graphs to visualize the association between getting a cardiovascular event (*Yes/No*) and two variables of your choice. Consider selecting one numerical and one categorical variable (it is not necessary to motivate the choice of these variables). Motivate your choice of graphs and describe what you observe [**5 points**].

(c) Estimate a logistic regression model for the probability of getting a cardiovascular event on the entire data set with `EVENT` as a dependent variable and the variables: `AGE`, `SEX`, `BMI`, `SYSTH`, `HDL`, `DIABETES`, `HISTCAR2`, `HOMOC`, `log(CREAT)`, `STENOSIS`, `IMT`, `SMOKING`, `ALCOHOL`, and `ALBUMIN` as predictors (include all these variables in the model, NO variable selection at this stage). Briefly describe your findings on the model estimated on the entire data set (what is the statistical significance and size of the effect of the variables on the outcome in terms of, e.g., odds ratios, use the appropriate test in case of categorical variables) [**5 points**].

(d) The data set SMART.csv contains missing values. Use the library `mice` to impute the data (generate 5 imputed data sets) and estimate the same logistic regression model as in (c). Pool the outcomes of the 5 imputed data sets and compare the pooled estimates with the outcome from (c) [**10 points**].

PART 2: (deadline: September 27, 17:00) [15 points]

In part 1, you estimated a logistic regression model on the entire data set. In this part, you will validate this logistic regression model. Again all variables should be included in the model (no variable selection at this stage).

(a) Assess and discuss the performance (predictive ability) of the logistic regression model using cross-validation. Use the AUC with an ROC curve as a performance measure [**5 points**].

(b) Assess and discuss the calibration of the model, i.e., the agreement between observed outcome and prediction, using a calibration plot [**5 points**].

(c) Compare the performance (predictive ability) of the logistic regression model with the performance of a linear discriminant analysis (LDA) model. Use the same predictors and estimate both models on the same data. When comparing the performances take explicitly into account the uncertainty in the estimates of the performance measure [**5 points**].

PART 3: (deadline: October 7, 17:00) [30 points]

In this part, you will apply various methods for variable selection and use several methods to validate models.

(a) Use *lasso* for variable selection to build a logistic regression model with (automatic) tuning for the probability of getting a cardiovascular event with all available predictor variables in the data set. Remember to specify the binomial family, otherwise, you will use linear regression. Use `log(CREAT)` instead of `CREAT` and remove `HISTCARD` and `HISTCAR2` to prevent multicollinearity. Tune the model with cross-validation or grid search, and validate the model with **training/test** validation. Describe your findings on this model (which variables were selected or most important) and the performance of the model (i.e., predictive ability with AUC and ROC curve) [**10 points**].

(b) Use *stepwise backward* variable selection to build a logistic regression model starting with all predictor variables (use the same predictors as in (a)) in the data set and validate that model using **cross-validation**. Describe your findings, i.e., the performance (AUC and ROC curve) of the final model (with confidence interval for the AUC) and the variables included in the final model [**10 points**].

(c) Validate the model from (b) using **bootstrap validation** and compare the obtained performance estimate (with confidence interval) with the validation results of question (b). For this question, you will write your **own** bootstrap validation function using the library `boot` [**10 points**].

PART 4: (deadline: October 21, 17:00) [30 points]

In this part, you will first develop a random forest model with tuning and validation. Next, you will build and validate a survival model using a Cox regression.

(a) Use *random forest* (including tuning) to develop and validate a prediction model based on all sensible predictor variables. Describe your findings (in terms of attribute importance) and assess the performance of the model (i.e., predictive ability with AUC and ROC curve with **training/test validation**) [**10 points**].

(b) In this data set, there are **two** associated endpoints: **`EVENT`** and **`TEVENT`**. So far, we have built classification models using the endpoint **`EVENT`**. Next to the presence of a cardiovascular event (**`EVENT`**), we know for all patients when the event occurred (**`TEVENT`**): the number of days since the event or the time a patient was in the cohort and censored at the end (but no event occurred). With these two endpoints combined you can build a survival model. Develop a *Cox regression* model using *stepwise backward* selection with the time outcome **`TEVENT`** combined with the presence of a cardiovascular event **`EVENT`**. Validate the model using **cross-validation.** Describe the final model in terms of the model performance and effect size / significance of the predictors of the final model [**10 points**].

(c) Write your own **bootstrap** function (using the library `boot`) to assess the stability of the stepwise backward selection combined with the *Cox regression* model in terms of which predictor variables are included in the model. If you bootstrap the entire procedure, you can assess how often a given predictor is included in the final model across all bootstrap samples [**10 points**].