RESULTS BY MODELS :

**WITH ALL THE PREDICTORS :**

Linear regression :

MSE = 0.1727088

Logistic regression :

MSE = 0.1727088

Cross Validation =

- for K = 10 : 0.1341752

- for leave-one-out (more precise) : 0.1319878

Polynomial regression + splines : *later in the tests*

Tree-based method :

«trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen »

BUT we’ll use it to have a general impression and view of the situation.

Classification tree :

Prunning tree :  
«  A smaller tree with fewer splits might lead to lower variance and better interpretation at the cost of a little bias.»

Size = 9

We won’t do it again after best subset selection : it is already « blurry » enough

**AFTER BEST SUBSET SELECTION :**

**Subset of predictor selected :**

BIC : 7 variables

*image*

Cp : 9 variables

*image*

Adjusted R square : 10 variables

*image*

*Therefore, the determination of which model of a given size is best must be made using only the training observations .*

MSE : We find the best model has 8 variables.

image

CROSS VALIDATION :

We find the best model has 11 variables, with cv = 0.1294388

However we can notice that the model with 7 variables has a cv = 0.1305488, which is very close to the previous one : difference = 0.00111.

Indeed, having 7 variables is more comfortables than having 11.

*Image*

Logistic Regression (with 11 variables):

MSE = 0. 1319878

Cross Validation =

- for K = 10 : 0. 1292328

- for leave-one-out : 0.1303879

GAM *(if you need infos : end of lecture 5, explain what it is…, how usefull can be… )*:

*using smoothing splines rather than natural splines*

Model2 : 11 values

Functions :

Linear for : sex, cp, trestbps, restecg, thalach, exang, slope, ca, thal

Quadratic : chol, oldpeak

MSE : 0.1121472

Image

plot

**NOT TESTED : EXPLAIN WHY**

- PCA : we have a response, we already have y. (chapt.7)

- bootstrap : not accurate .. (maybe explain why)

we don’t have many datas, BUT we’re working on something specific, where we’re not sure how uniques our datas are, so not sure we can generalize like that.

- technique tuto 9 -> last slide of chapter 9

**WHEN WE HAVE TO TALK ABOUT THE RESULTS WHICH ALREADY EXIST :**

On the kaggle page, there are analysis : explain we want to focus on the essential : heart disease or no -> ? and really find the best way blah blah blah