CENTRE DE RECHERCHE PUBLIC

Applications of variable selection methods to biomedical problems

International Year of Statistics Luxembourg

Olivier Collignon, PhD
olivier.collignon@crp-sante.lu

Context

Data

Outcome to describe and/or predict :

$$y \in \mathbb{R}, \{0,1\}, \ldots$$

Vector of predictors :

$$x = (x_1, \ldots, x_p)$$

We want to use the vector of predictors to model the outcome.

Context

Which are the most relevant variables to describe and/or predict *y* ?

Examples :

(1) Biomarkers and SNP that help predict blood drug concentration in HIV patients ?

(2) Risk factors that favor an occurrence of myocardial infarction ?

(3) Risk factors and tumor descriptors (eg : TNM) that best describe cancer survival ?

Context

Classic models

Linear regression :

$$y|x = \beta_0 + \beta x + \epsilon, \epsilon \cong \mathrm{N}(0, \sigma^2)$$

Logistic regression

$$\mathrm{logit}\big(P(y = 1|x)\big) = \beta_0 + \beta x$$

Cox proportional hazards regression

$$h(t|x) = h_0(t)\exp(\beta x)$$

Variable selection

Why to eliminate variables ?

_ More precise description of the phenomenon ;

_ Costs ;

_ Robustness, errors… ;

_ Some methods are not available anymore when $p > n$;

_ …

Variable selection

Whitehead's rule to avoid overfitting

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$
Failure (survival) time	number of failures

Enter m/20 to m/10 variables in the model

Variable selection

3 frequently used methods :

1. Filter based methods using univariate testing ($p < 0,05$) ;

2. Automated selection ;

3. Penalization techniques

Variable selection

1. Filter based methods using univariate testing ($p < 0,05$)

Useful to evaluate the effect on each predictor independantly, but:

_ does not account for the multi-variables aspect ;
_ some variable may become significant only in the presence of other ones ;
_ criterion based on inference instead of prediction

Useful to convince oneself but obsolete

Variable selection

2. Automated variable selection (forward)

The number of combination of variables among p is about 2^p , which is harldy handled for p>15 ;

Start the model with the intercept only ;

a) Add to the model the most significant variable (if any), according to appropriate test ;

b) Add the variable that brings the more complementary information to preceding model (if any among all remaining ones), according to appropriate test ;

c) …

d) The algorithm stops when no remaining variable is significant or if every variable have already been entered.

The algorithm also exists in backward / stepwise mode.

Variable selection

Limitations :

1. The effect of the predictors on the outcome may be be overestimated (Chatfield,1995) ;

2. The more candidate predictors, the more <<noisy>> variables retained in the final model (Derksen et Keselman, 1992) ;

3. Two highly correlated predictors are unlikely to be both selected together even if their joint effect is useful ;

4. ….

If parcimony is needed : bootstrap !

3. Penalization

Penalization techniques avoid parameter estimates being artificially large (Harrell, 2001).
By denoting *L(β)* the model likelyhood, we want to maximize the quantity :

$$L(\beta) - \frac{\lambda}{2} pen(\beta)$$

Where *pen* is a penalization funtion to choose between (among others) :

$$pen(\beta) = \begin{cases} ||\beta||_1 & L^1 \text{ criterion (LASSO like)} \\ ||\beta||_2^2 & L^2 \text{ criterion (Ridge regression like)} \\ (1-t)||\beta||_1 + t||\beta||_2^2 & \text{Elastic Net (EN)} \end{cases}$$

The optimal *λ* value is obtained by minimizing the Akaike Information Criterion (AIC) of the model within a grid of values (same thing for *t* in Elastic Net)

Application

LUCKY Registry

n = **135** STEMI patients who signed a written informed consent.

The New-York Heart Association score measures the limitation of patients' physical activity.

NYHA	symptoms
I	no limitation is experienced in any activities; there are no symptoms from ordinary activities.
II	slight, mild limitation of activity; the patient is comfortable at rest or with mild exertion.
III	marked limitation of any activity; the patient is comfortable only at rest.
IV	any physical activity brings on discomfort and symptoms occur at rest.

Application

25 variables (mesured by research nurses via a form) and divided into several subgroups :

1. **Identification** : *age, sex, weight* ;
2. **Biomarkers** : *MMP9, TIMP1, WBC, hsCRP* ;
3. **History** : *infarct, PTCA, diabetes, hypertension, cholesterol, family history, smoking habits* ;
4. **Medication** : *aspirin, plavix, b blockers, ca2 antagonist, nitroglycerin, CEI, statines, angiotensin II receptor antagonist, diuretics, heparin, reopro.*

Which features in the registry are linked to an elevated NYHA score at the end of follow-up ?

Application

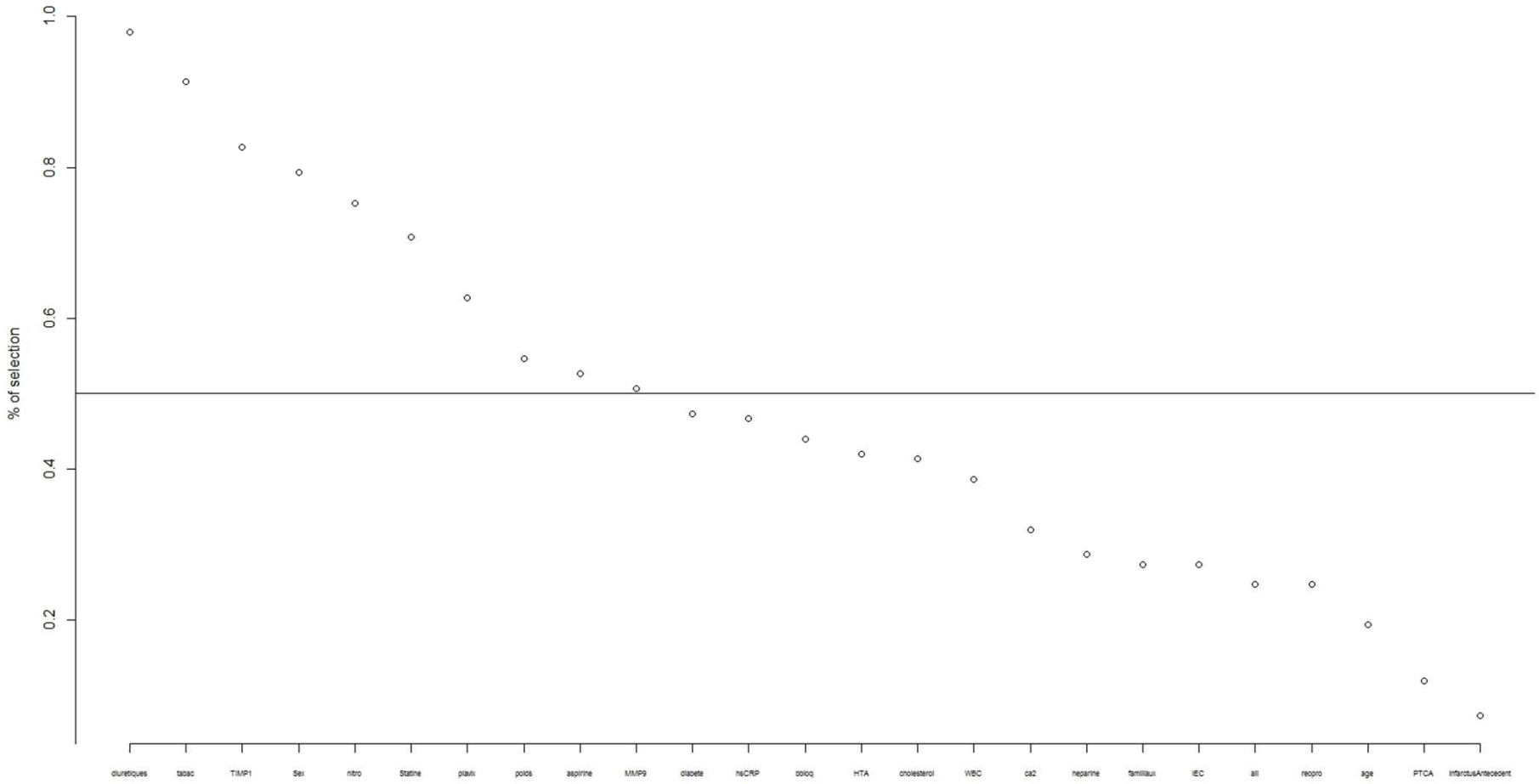
- Use L^1 penalty because it guarantees sparsity (*i.e.* few predictors)
- Force *t=0* in Elatic Net settings

```
fit$glmnet.fit$beta[,n.lambda.min]
        age       poids        MMP9       TIMP1         WBC
 0.00000000  0.00000000  0.00000000  0.01011795  0.00000000
        HTA cholesterol   familiaux       tabac    aspirine
 0.00000000  0.00000000  0.00000000 -0.49353160  0.06441077
    Statine         all  diuretiques    heparine     reopro
 0.05785930  0.00000000  0.68004621  0.00000000  0.00000000
      hsCRP         Sex infarctusAntecedent        PTCA    diabete
 0.00000000 -0.39825333  0.00000000  0.00000000  0.00000000
     plavix       bbloq         ca2       nitro         IEC
 0.16073490  0.00000000  0.00000000 -0.13239325  0.00000000
```

Leads to 94% of correct classification rate using logistic regression

Application

n=150 bootstrap samples



References

1. Frank E Harrell, *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis,* Springer Verlag, New-York,2001

2. Willy Sauerbrei, Patrick Royston, Harald Binder, *Selection of important variables and determination of functional form for continuous predictors in multivariable model building*, Statistics in Medicine, 26, 30, pp 5512–5528, 2007