Reminder TD$_5$

Marion Brouard, Pauline Leveneur

November 19, 2025

- **Objective of the review:** Understanding the concepts of
  1. Non-causal predictions
  2. Variable selection
  3. Penalized regressions

1 Non-causal predictions

We seek to predict as well as possible a variable Y_{n+1} from X_{n+1} and a sample *iid* $(Y_i, X_i)_{i=1,\dots,n}$ of size n. **We are not trying to find out whether the coefficients of X represent a causal effect**: we are only interested in the prediction. For example, we observe the wage of an individual for the years 1 to n and the education, age and sex for the years 1 to $n+1$ and we want to predict the wage of the individual for the year $n+1$.

To get an idea of the quality of the made prediction, we place ourselves on the training subset $\mathcal{T}$ of size n_T, we wish to predict as well as possible $\hat{Y}_{n_T+1}$ from X_{n_T+1}. We are therefore looking for the best parameter $\beta(\mathcal{T})$ which solves

$$\underset{\beta(\mathcal{T})}{min} \left[(\underbrace{Y_{n_T+1}}_{\text{observed}} - \underbrace{X'_{n_T+1}\beta(\mathcal{T})}_{\text{predited}})^2 \right]$$

<u>In practice:</u>

1. We separate the data into a *train* sample (on which we train the model, for example 70% of the observations) and a *test* sample (on which we test the model, for example 30% of the observations).
2. The model is trained on the *train* sample → the $\hat{\beta}$ are recovered
3. The $\hat{Y}$ of the *test* sample are predicted from the X of the *test* sample and the $\hat{\beta}$ obtained in step 2.
4. The prediction error is calculated by comparing the $\hat{Y}$ predicted in step 3 and the Y observed in the *test* sample. For example, we can use the mean square error: MSE $= \frac{1}{n}\sum_i(\hat{Y}_i - Y_i)^2$

2 Variable selection

2.1 Overfitting

The number of explanatory variables in a model determines the number of parameters to be estimated. However, estimating (too) many parameters can lead to overfitting and a lack of precision. Overfitting means that the trained model fits the particular data used too precisely. By adding too many variables, we try too hard to fit the sample data, and the model thus becomes difficult to generalize. In order to avoid a model that is over-calibrated by the database being studied, we may need to select the variables to be included in the model. This can be done using information criteria.

2.2 Criteria for variable selection

Several information criteria can be used to select variables.
An information criterion is a quantity that measures the goodness of fit of a statistical model. The AIC (*Akaike information criterion*) and the BIC (*Bayesian information criterion*) are based on the log likelihood. It is also possible to use the adjusted R^2, which is calculated from the explained variance.
When denoting k the number of explanatory variables (regressors), n the sample size, and $widehat\mathcal{L}_n$ the empirical log-likelihood evaluated at its maximum:

$$\begin{aligned}
R^2_{\text{aj}} &= 1-(1-R^2)\frac{n-1}{n-k}, \text{ cf TD2}\\
AIC &= -2\widehat{\mathcal{L}}_n + 2\times k, \text{ cf chapter 3, slide 24}\\
BIC &= -2\widehat{\mathcal{L}}_n + ln(n)\times k, \text{ cf chapter 3, slide 24}
\end{aligned}$$

These three criteria penalize the number of regressors.
In the case of the adjusted R^2, we check whether the addition of a variable, which leads to a mechanical increase in the R^2 (R^2 weakly increasing in k), is compensated for by the penalty of the number of variable k ($-\frac{n-1}{n-k}$ decreasing in k).
In the case of AIC and BIC, for the addition of a new variable to be informative, the log-likelihood must increase by at least the increase in the penalty (i.e. $2timesk$ for AIC and $ln(n)timesk$ for BIC).
A good prediction quality corresponds to a high R^2_{adj} and a low AIC and BIC.

<u>In practice:</u>

1. I test several regressions (including different explanatory variables) and calculate the AIC, BIC, R^2_{adj} for each one.

2. I keep the one with the smallest AIC or the smallest BIC or the largest R^2_{adj}.
(Warning: these criteria may not select the same models: BIC penalizes the addition of new variables more heavily)

3. I predict using only the selected variables.

3 Penalized regressions

An alternative to the information criterion, but with the same purpose of avoiding overfitting, is to use penalized regressions. Unlike information criteria, which simply calculate a criterion for each regression performed (the regression with the lowest criterion is then chosen), penalized regressions allow all the coefficients to be estimated, taking into account the penalties. Two types of penalized regression are presented here: Lasso regression and Ridge regression.

3.1 Introduction

The Lasso and Ridge estimators are obtained by the following minimization programs:

$$\hat{\beta}_{LASSO} = \arg\min_{b \in \mathbb{R}^k} \quad \underbrace{\sum_i (Y_i - X_i' b)^2}_{\substack{\text{sum of squared residuals} \\ \text{(like OLS)}}} \quad + \quad \underbrace{\lambda ||b||_1}_{\substack{\text{penalty term} \\ \text{or regularization}}}$$

$$\hat{\beta}_{RIDGE} = \arg\min_{b \in \mathbb{R}^k} \sum_i (Y_i - X_i' b)^2 + \lambda ||b||_2^2$$

λ is a positive hyperparameter to choose from.

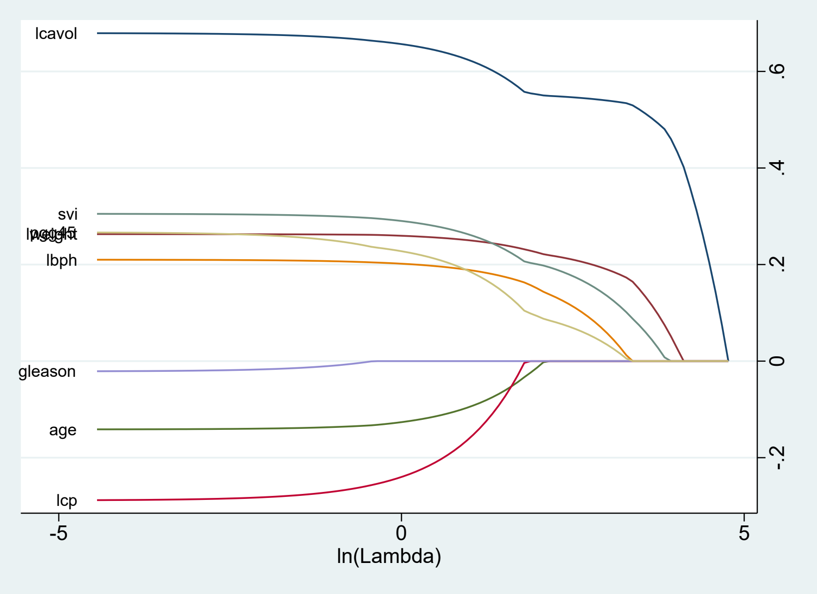
Unlike OLS, Lasso and Ridge regressions add a penalty for the size (in absolute value) of the coefficients. **The idea is that we want to keep the variables that are really of interest (second term) while ensuring that we obtain a small error with them (first term).**

In the case of Lasso, several coefficients can be estimated at 0 ($\hat{\beta}^j_{LASSO} = 0$), which amounts to an automatic exclusion of those variables from the regression.

These regressions take into account the 1 or 2 norm of the coefficients of the variables, which necessarily depend on the scale of the variables. Thus, if our variable Y is small, but one of the X^j is large, then its coefficient will probably be very low, as the Lasso would probably force it to be equal to (or very close to) 0. Even though a small variable would have a larger coefficient, simply because of the size of the variables. Let's take an example: we want to predict age at death using a Lasso regression. The 'income' variable is *a priori* a very relevant variable for predicting age at death. However, one extra euro of income will have only a minimal impact on your age at death, so the Lasso regression will probably set the value of the coefficient at 0. To prevent this size effect, we need to **standardize the variables beforehand**. The result is that there is no longer a constant in the model.

3.2 Example

The following graph plots the estimated coefficients according to the value of lambda.

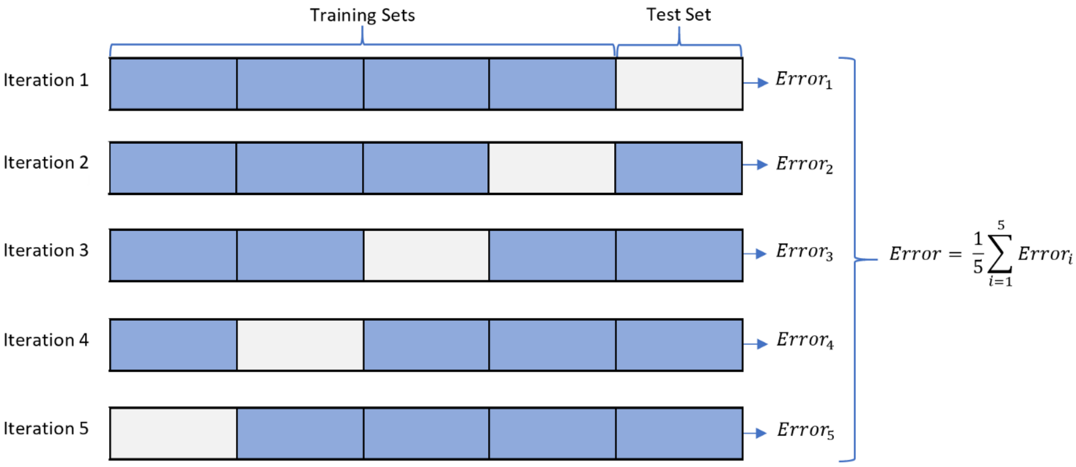


As the value of lambda increases, the coefficients tend towards 0. This means that the penalty of the norm is higher and higher, and that the penalty term becomes more and more important in the maximization program. **That is, the larger lambda is, the more we seek a parsimonious model with few variables and low coefficients**.

3.3 Lasso in practice

3.3.1 Choice of lambda by cross-validation

The principle of **cross-validation** is to keep the same data set but to use several partitions of the *train* and *test* samples.
NB: Cross-validation is used to select parameters. It is not specific to Lasso.



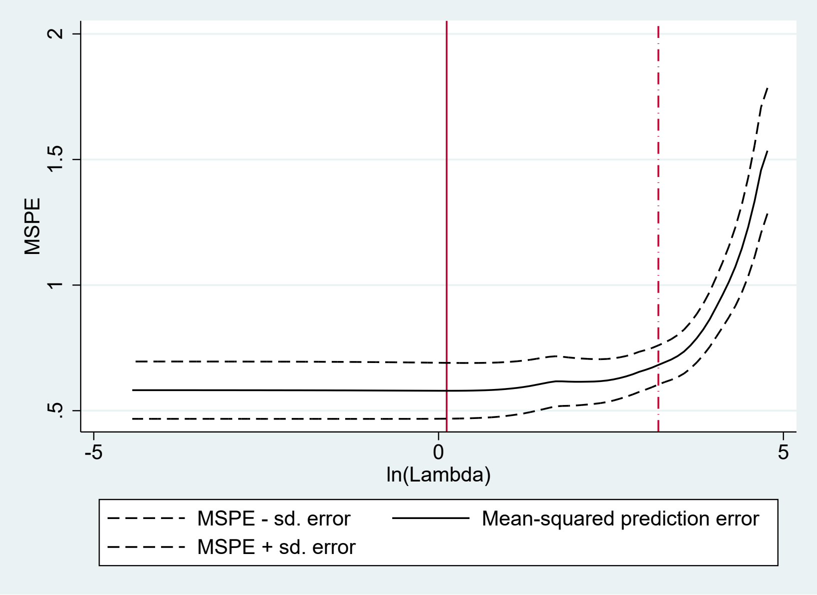
<u>In practice:</u>

1. I take a value of *lambda*.

2. For this λ, I calculate the error on each partition ('iteration' on the illustration above) and then I average the different partitions (mean square error, MSE)

3. Repeat steps 1. and 2. for each candidate lambda value

4. Obtain the following graph:
   On the ordinate: the mean square error (average calculated over all the partitions).
   x-axis: the different ln(lambda) values being tested.
   We then choose the λ which minimises the MSE (straight red line on the graph) or else the largest λ such that the MSE does not start to sharply increase (dotted red line on the graph).



3.3.2 Estimation

- **Lasso**

Lasso estimation is carried out using the optimal *ambda* (from cross-validation).
The estimation will select the variables: some coefficients are estimated at 0 at this stage ($\hat{\beta}^j_{LASSO} = 0$), so these variables will have to be deleted in the final model.

NB: if $lambda > 0$ all the coefficients are biased towards 0.

- **Post-Lasso**

Post-Lasso estimation is a two-stage procedure. The first part consists of estimating a LASSO model (with the optimal λ) which selects the variables. The second step consists of estimating an OLS which includes only the variables whose coefficients are different from 0 in the first step, i.e. a linear regression (without penalty) on the variables retained by the Lasso.

To sum up:

1. **Cross validation** $\rightarrow \lambda^\star$

2. **Lasso** (penalized regression) with $\lambda^\star \rightarrow \hat{\beta}_{LASSO}$

3. **Post-Lasso** : OLS (unpenalized regression) on covariates selected at step 2, *ie* X_j such that $\hat{\beta}^j_{LASSO} \neq 0$