Master in Financial Engineering (EPFL)
Financial Econometrics

Part II: Machine learning and Asset Pricing
Lecture 2: Selection of variables—Subset selection methods

Florian Pelgrin

EDHEC Business School

February - June 2019

1. Introduction

2. Subset selection
   - Best subset selection
   - Forward stepwise selection
   - Backward stepwise selection
   - Choosing the optimal model

1. Introduction

Problem

- Consider the standard (generic) linear model (with n observations)

$$y = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + u$$

where p is the number of (non-constant) explanatory variables.

- What could be done in the following cases?
  - p is large but $p < n$
  - $p >> n$.

Briefly speaking,

- If $\mathbf{n} >> \mathbf{p}$, least squares estimates tend generally to perform well;
- If n **is not much larger than** p, there can be a lot of variability, resulting in overfitting and poor prediction accuracy;
- If $\mathbf{p} > \mathbf{n}$, then least squares estimates are no longer unique and the variance is *infinite*. Especially, one needs to rely on constraints and shrinkage methods.

Solutions

- **Subset selection**: Select a subset of the p variables for (each) possible combinations of the p predictors and fit a model on the reduced set of variables;
- **Shrinkage**: Fit a model with all variables but use a shrinkage (regularization) procedure (e.g., some (all) coefficients are shrunken towards zero);
- **Dimension reduction**: Project the p explanatory variables into a m-dimensional subspace with $m < p$ and fit a model with m projections or linear combinations of initial variables.

Examples

✓ Example 1: Credit rating

- (Small!) Set of predictors: "Income", "Limit", "Rating", "Cards", "Age", "Education", "Gender", "Student","Married","Ethnicity"
- Output variable: "Balance";

✓ Example 2: UK CPI inflation forecasting

- (Small) set of *quarterly* macrovariables for the UK between 1988Q1 and 2015Q4;
- Use a lead-lag model in which the output/target variable, CPI inflation, leads changes or the level of other features/variables by two years;

Among other questions, How to select the best specification? In which sense?

Main objectives

✓ Overview of main challenges:

- Selection methods
- Shrinkage methods
- Reduction methods

✓ Review some applications and the programming.

2. Subset selection

2.1. Best subset selection

Algorithm: Best subset selection

1. Let $\mathcal{M}_0$ denote the *null model* (with no predictors).
2. For $k = 1, \cdots, p$:
   - ✓ Fit all $\binom{k}{p}$ models with exactly k predictors;
   - ✓ Pick the best model, denoted $\mathcal{M}_k$, among these $\binom{k}{p}$ models using some goodness-of-fit measures (e.g., R^2);
3. Select a single best model using the $p+1$ options $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using cross-validated prediction error, AIC, BIC or adjusted R^2.

Issues:

- ✓ If $p > n$, it can be used only up to n variables (with poor performances as p gets closer to n);
- ✓ The curse of dimensionality: If $p = 10$ (resp., 20), it requires approximately 1'000 possible models (resp., over one millon possible models). More generally, it involves fitting 2^p models. In practise, it becomes computationally infeasible for $p > 40$.
- ✓ It works only for least squares linear regressions;
- ✓ There is a risk of overfitting for large p and thus of high variance of the coefficient estimates.

Note: Some possible models can be eliminated with branch-and-bound techniques as long as p is not too large.

Remark: The best subset selection model can be written as the following constrained minimization problem:

$$\min_{\beta_0,\cdots,\beta_p} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 \right\}$$

$$\text{s.t.} \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$$

where $I(\beta_j \neq 0)$ is an indicator variable that takes on a value of 1 if $\beta_j \neq 0$ and equals zero otherwise. After demeaning variables, this is also equivalent to:

$$\widehat{\beta}(\lambda) = \underset{\beta}{\texttt{argmin}} \left(\|y - X\beta\| + \lambda \|\beta\|_0^0 \right)$$

where $\beta = (\beta_1, \cdots, \beta_p)^\top$ and $\|\beta\|_0^0 = \sum_{j=1}^{p} I(\beta_j \neq 0)$ is the ℓ_0-penalty (or the ℓ_0 analogue of a norm).

2.2. Forward stepwise selection

Algorithm: Forward stepwise selection

1. Let $\mathcal{M}_0$ denote the *null model* (with no predictors).

2. For $k = 0, \cdots, p-1$:

   - ✓ Consider all $p-k$ models that augments $\mathcal{M}_k$ with one additional explanatory variable;
   - ✓ Choose the *best* model, denoted $\mathcal{M}_{k+1}$, among these $p-k$ models using some goodness-of-fit measures (e.g., R^2);

3. Select a single best model using the $p+1$ options $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using cross-validated prediction error, AIC, BIC or adjusted R^2.

In practise,

- ✓ Step 1: Start with $\mathcal{M}_0$

$$y_i = \beta_0 + u_i$$

- ✓ Step 2: Loop $k = 0, \cdots, p-1$
  - $k = 0$: estimate each (p) model and use a goodness-of-fit measure (e.g, $\bar{R}^2$)

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + u_{i,1} \\ y_i &= \beta_0 + \beta_2 x_{i,2} + u_{i,2} \\ &\vdots \\ y_i &= \beta_0 + \beta_p x_{i,p} + u_{i,p} \end{aligned}$$

Choose the best specification, denoted $\mathcal{M}_1$, using the goodness-of-fit measure

In practise (cont'd),

- ✓ Step 2 (cont'd)
  - $k = 1$: Consider all $(p-1)$ specifications and use a goodness-of-fit measure

$$y_i = \underbrace{\beta_0 + \beta_\ell x_{i,\ell}}_{\mathcal{M}_1} + \beta_j x_{i,j} + u_{i,j}$$

    where $j \in \{1, \cdots, p\}$ and $j \neq \ell$.
    Choose the best specification, denoted $\mathcal{M}_2$, using the goodness-of-fit measure.

  - ...
- ✓ Step 3: Choose among the collection of (adjusted) models $\mathcal{M}_0, \cdots, \mathcal{M}_p$.

Remarks

- ✓ Forward stepwise selection methods involves fitting the null model and $p-k$ models for each iteration $k = 0, \cdots, p-1$, i.e. $1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$ models.

  Example: if $p = 20$, the best subset selection method (resp., **forward stepwise selection**) involves 1'048'576 models (resp., **211 models**).

- ✓ It is not guaranteed to find the best possible model out of all 2^p models, i.e. the selection of variables can be different.

- ✓ It can be applied when $p > n$ but only by considering the sub-models $\mathcal{M}_0, \cdots, \mathcal{M}_{n-1}$.

Remarks (cont'd)

- Best subset selection method and forward stepwise selection method can lead to different choices, especially when imposing some constraints on the number of features/explanatory variables

Variable selection with credit data

Nb of features	Best subset	Forward stepwise
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	**cards**, income, student, limit	rating, income, student, limit

2.3. Backward stepwise selection

Algorithm: Backward stepwise selection

1. Let $\mathcal{M}_p$ denote the *full model* (with all features/predictors).
2. For $k = p, \cdots, 1$:
   - ✓ Consider all k models that contain *all but one* of the variables in $\mathcal{M}_k$ for a total of $k-1$ variables;
   - ✓ Choose the *best* model, denoted $\mathcal{M}_{k-1}$, among these k models using some goodness-of-fit measures (e.g., R^2);
3. Select a single best model using the $p+1$ options $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using cross-validated prediction error, AIC, BIC or adjusted R^2.

In practise,

- ✓ Step 1: Start with $\mathcal{M}_p$

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j + u_i$$

- ✓ Step 2: Loop $k = p, \cdots, 1$
  - $k = 1$: estimate each (p) model and use a goodness-of-fit measure (e.g, $\bar{R}^2$)

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=2}^{p} x_{i,j}\beta_j + u_{i,1} \\ y_i &= \beta_0 + \sum_{j\neq 2}^{p} x_{i,j}\beta_j + u_{i,2} \\ &\vdots \\ y_i &= \beta_0 + \sum_{j=1}^{p-1} x_{i,j}\beta_j + u_{i,p} \end{aligned}$$

Choose the best specification, denoted $\mathcal{M}_{p-1}$, using the goodness-of-fit measure

In practise (cont'd),

✓ Step 2 (cont'd)

- $k = p - 1$: Consider all $(p-1)$ specifications and use a goodness-of-fit measure

$$y_i = \underbrace{\beta_0 + \sum_{j \neq \ell} x_{i,j}\beta_j}_{\mathcal{M}_p} - \beta_k x_{i,k} + u_{i,j}$$

where $k \in \{1, \cdots, p\} \setminus \{\ell\}$.
Choose the best specification, denoted $\mathcal{M}_{p-1}$, using the goodness-of-fit measure.

- ...

✓ Step 3: Choose among the collection of (adjusted) models $\mathcal{M}_0, \cdots, \mathcal{M}_p$.

Remarks

- ✓ Backward stepwise selection methods involves fitting the full model and k models for each iteration $k = 1, \cdots, p$, i.e. $1 + \sum_{k=1}^{p} k = 1 + \frac{p(p+1)}{2}$ models.

  Example: if $p = 20$, the backward stepwise selection involves 211 models as in the case of the forward stepwise selection.

- ✓ It is not guaranteed to find the best possible model out of all 2^p models, i.e. the selection of variables can be different.

- ✓ It cannot be applied when $p > n$, i.e. the only viable selection method in this context is the forward stepwise approach. At the same time hybrid versions of forward and backward stepwise selection can be implemented.

2.4. Choosing the optimal model

The optimal model is generally chosen using two common approaches:

1. *Indirect* estimation of the *test error* by adjusting the *training error* for the model size (information criteria);

2. Direct estimation of the test error by making use of a validation set or a cross-validation approach.

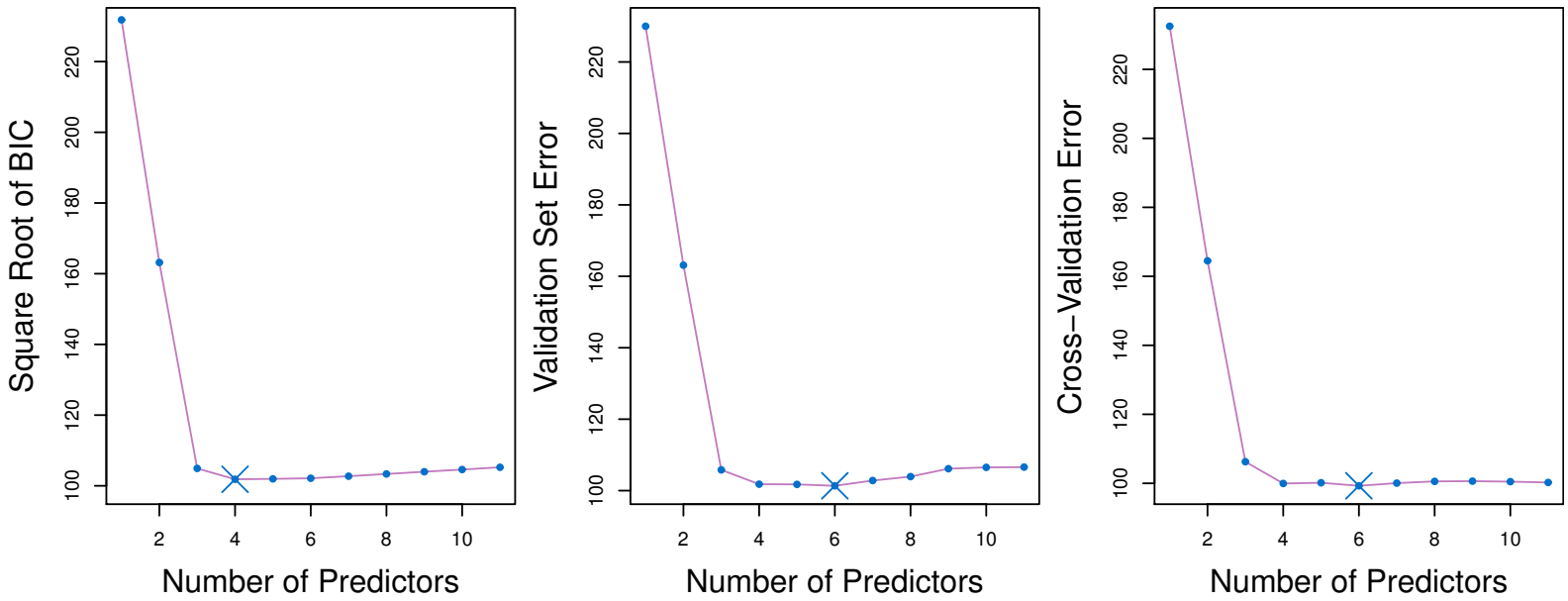

Figure: Credit data: Validation and Cross-validation

Note: Left panel—(square root of) BIC, Center panel—Validation set errors, Right-panel: 10-fold Cross-validation errors