Model Selection

- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- Adding new variables automatically reduces the RSS.
- But it does not improve either interpretablility nor prediction on a test set.
- Pblm : How to select a subset of variables which offers good prediction and avoids overfitting ?

Model Comparison Criteria : AIC, BIC, C_p

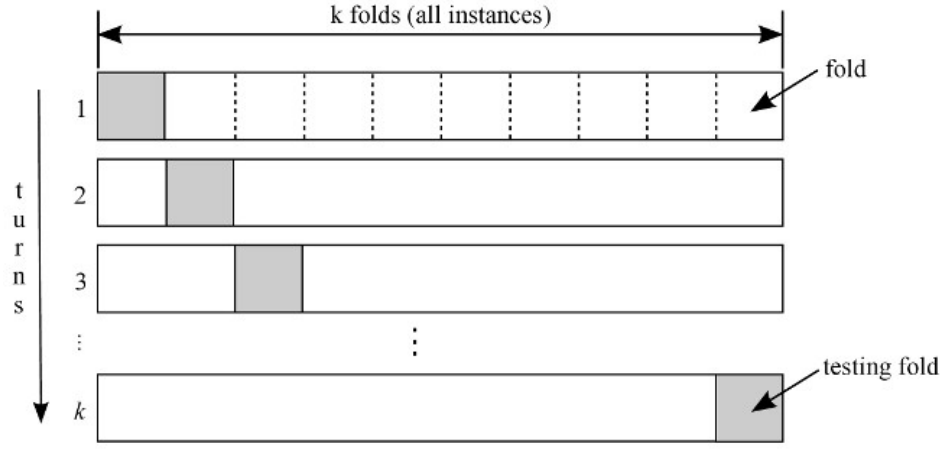
- Different criteria can be used.
- Akaike information criterion, $AIC := -2 \log L + 2d$ with L the maximized likelihood and d the number of variables.
- Bayesian information criterion $BIC = n^{-1}(RSS + \log(n) d\sigma^2)$.
- Mallow's $C_p := n^{-1}(RSS + 2d\sigma^2)$ which is equivalent to AIC for the Gaussian linear regression.
- Given one of these criteria, a model is better than another if its criterion value is smaller.

Model Comparison Criteria : Cross Validation

- The sample is randomly divided between a training and test sets (*e.g.*, $80\% - 20\%$ fold).
- The parameter is estimated on the training set. MSE is computed on the test set.
- Model with smallest MSE is selected.
- Advantage : provide a direct measure of test error and does not require to estimate σ.
- Robust indicator that can be used even if the number of degrees of freedom is unknown.

Model Comparison Criteria : Cross Validation Detailed

1. Data randomly shuffled.
2. The sample is divided between test and training set, turn by turn.
3. At each turn, the model is estimated on the training set and MSE computed on the test set.
4. The result is obtained by averaging the MSEs obtained at each turn.



Selection Methods : Subset Selection

- Idea : Given a criterion, calculate it for every possible model and select the model with the best criterion value.
- Advantage : Every possible model is covered.
- Drawbacks : Computationally inefficient ($\sim 2^p$).

Selection Methods : Forward Selection

- Idea : Start from the empty subset and then add variables one by one until reaching the model with all variables. Select the model with the smallest criterion value.
- Advantage : More efficient than subset selection ($\sim p^2$).
- Drawbacks : No guarantee to select the best model (relatively to the chosen criterion) among all possible models. Polynomial complexity $=>$ cannot scale to some large datasets.

Selection Methods : Forward Selection detailed

1. Start with the empty model M_0 and the sets $\mathcal{A} = \emptyset$, $\mathcal{I} = \{1, \cdots, p\}$
2. For $k = 1, \cdots, p-1$
   - 2.1 $i = \arg\min_{j \in \mathcal{I}} RSS_{\mathcal{A} \cup \{j\}}$
   - 2.2 Add the variable which minimize RSS : $\mathcal{A} \leftarrow \mathcal{A} \cup \{i\}$, $\mathcal{I} \leftarrow \mathcal{I} \setminus \{i\}$
   - 2.3 M_k is the model with variables of set $\mathcal{A}$
3. M_p is the full model
4. Select the best (relatively to the chosen criterion) model among $M_0, \cdots, M_p$

Selection Methods : Forward Selection example

Full model : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

Model	Equation	AIC	BIC
M_0	$Y = \beta_0 + \epsilon$	160	180
M_1	$Y = \beta_0 + \beta_3 X_3 + \epsilon$	150	**165**
M_2	$Y = \beta_0 + \beta_3 X_3 + \beta_1 X_1 + \epsilon$	**145**	170
M_3	$Y = \beta_0 + \beta_3 X_3 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$	152	173

Selection Methods : Backward Selection

- Idea : Start from the full model and then remove variables one by one until reaching the empty model. Select the model with the smallest criterion value.
- Advantage : More efficient than subset selection ($\sim p^2$).
- Drawbacks : No guarantee to select the best model (relatively to the chosen criterion) among all possible models. Polynomial complexity $=>$ cannot scale to some large datasets. Inefficient when p is large because it starts from the full subset.

Selection Methods : Backward Selection detailed

1. Start with the full model M_p and the sets $\mathcal{A} = \{1, \cdots, p\}$
2. For $k = p - 1, \cdots, 1$
   - 2.1 $i = \arg\min_{j \in \mathcal{A}} RSS_{\mathcal{A} \setminus \{j\}}$
   - 2.2 Remove the variable which minimize RSS : $\mathcal{A} \leftarrow \mathcal{A} \setminus \{i\}$
   - 2.3 M_k is the model with variables of set $\mathcal{A}$
3. M_0 is the empty model
4. Select the best (relatively to the chosen criterion) model among $M_0, \cdots, M_p$

Selection Methods : Selection with Stopping Criterion

- Forward and Backward procedures also exist in a stopping criterion version :
  - Forward : If the newly added variable is not significant enough (partial F-test) the algorithm stops
  - Backward : If the variable to remove is too significant (partial F-test) the algorithm stops
- Combining the two (forward then backward after the forward algorithm stops) is known as stepwise selection

LASSO regression

- Idea : reformulate the least square problem so that its minimization favors sparse solution.
- LASSO regression :

$$\hat{X}_{\text{LASSO}} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \| Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p \|_2^2 + \lambda \|\beta\|_1$$

  with $\lambda > 0$ and $\|\beta\|_1 := \sum_{i=0}^{p} |\beta_i|$.
- Advantage : It is a convex problem that can be solved efficiently and scaled to large datasets. The regularization term λ favors sparse solutions.

LASSO regression : Choosing λ

- One needs to choose the tuning parameter λ.
- This is done via cross validation : the MSE is computed for each value of λ and one choses the value for which it is minimum.
- The model is then re-fit on the entire dataset for the chosen value of λ.

LASSO regression : Choosing λ