Model Selection Algorithm and Guideline

Model selection algorithm

- “Best” subsets algorithms
  - Provide the best subsets according to the specified criterion and identify several good subsets for each possible number of X variables to give the additional information.
  - Use when the potential X variables is relatively small, <30.
- Stepwise regression methods
  - Develops the best subsets sequentially.
  - Contain both forward selection and backward elimination.
  - Only a single regression model is identified by the stepwise regression method. The last model Could be suboptimal.
- When the pool of X variables is very large, we should use the subset identified by the stepwise search procedures as a starting point. One may treat the selected number of X variables in the regression model as the right subset size and then use the “best” subsets procedures.

Case Study: Surgical Unit Example (8 X variables, $p = 9$)

A hospital surgical unit was interested in predicting survival in patient undergoing a particular type of liver operation. A random number of 108 patients was available for analysis, but we only study (n=)54. For each patient record, the following information was extracted (data: surgery.csv):

Potential predictors include,

- Blood clotting score (X_1, blood)
- A prognostic index (X_2, prog)
- Enzyme function test (X_3, enz)
- Liver function test (X_4, liver)
- Age (X_5, age)
- Gender ($X_6, gender\ 0 = male, 1 = female$)
- History of alcohol use (3 levels 2 indicator variables X_7 and X_8)

Alcohol Use	X7	X8
None	0	0
Moderate	1	0
Severe	0	1

The response variable is survival time in days (Y, surv)

“Best” subsets algorithms

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)

X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

```
library(ALSM)
bs<-BestSub(sur[,1:8], sur$lny, num=1)
```

```
  p 1 2 3 4 5 6 7 8     SSEp        r2    r2.adj         Cp      AICp       SBCp   PRESSp
1 2 0 0 1 0 0 0 0 0 7.331575 0.4275662 0.4165579 117.409441 -103.8269  -99.84889 8.326716
2 3 0 1 1 0 0 0 0 0 4.312491 0.6632899 0.6500855  50.471575 -130.4833 -124.51634 5.065339
3 4 0 1 1 0 0 0 0 1 2.842883 0.7780337 0.7647157  18.914496 -150.9849 -143.02899 3.469403
4 5 1 1 1 0 0 0 0 1 2.178799 0.8298840 0.8159970   5.750774 -163.3514 -153.40643 2.737771
5 6 1 1 1 0 0 1 0 1 2.082008 0.8374413 0.8205081   5.540639 -163.8052 -151.87127 2.782713
6 7 1 1 1 0 1 1 0 1 2.005225 0.8434363 0.8234494   5.787389 -163.8343 -149.91140 2.772325
7 8 1 1 1 0 1 1 1 1 1.972032 0.8460279 0.8225974   7.029455 -162.7356 -146.82378 2.808705
8 9 1 1 1 1 1 1 1 1 1.970742 0.8461286 0.8187737   9.000000 -160.7710 -142.87013 2.931232
```

Q: Among these 8 subsets, which is identified as the best under each criterion.

SSEp 8 R2 8 R2.adj 6 Cp 5 , AICp 6 ,SBC 4 , PRESSp 4

“Best” subsets algorithms

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)

X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

```
library(ALSM)
bs<-BestSub(sur[,1:8], sur$lny, num=3)
```

	p	1	2	3	4	5	6	7	8	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
1	2	0	0	1	0	0	0	0	0	7.331575	0.4275662	0.4165579	117.409441	-103.82686	-99.84889	8.326716
1	2	0	0	0	1	0	0	0	0	7.408731	0.4215420	0.4104178	119.171240	-103.26154	-99.28357	8.024956
1	2	0	1	0	0	0	0	0	0	9.979182	0.2208467	0.2058629	177.865004	-87.17808	-83.20011	10.743872
2	3	0	1	1	0	0	0	0	0	4.312491	0.6632899	0.6500855	50.471575	-130.48329	-124.51634	5.065339
2	3	0	0	1	1	0	0	0	0	5.129702	0.5994837	0.5837772	69.131808	-121.11257	-115.14561	6.120508
2	3	1	0	1	0	0	0	0	0	5.780964	0.5486346	0.5309340	84.002738	-114.65834	-108.69138	6.987582
3	4	0	1	1	0	0	0	0	1	2.842883	0.7780337	0.7647157	18.914496	-150.98493	-143.02899	3.469403
3	4	1	1	1	0	0	0	0	0	3.108539	0.7572918	0.7427294	24.980500	-146.16088	-138.20494	3.914240
3	4	0	1	1	1	0	0	0	0	3.614131	0.7178164	0.7008853	36.525190	-138.02317	-130.06723	4.596928
4	5	1	1	1	0	0	0	0	1	2.178799	0.8298840	0.8159970	5.750774	-163.35135	-153.40643	2.737771
4	5	0	1	1	1	0	0	0	1	2.376584	0.8144413	0.7992937	10.267014	-158.65926	-148.71434	3.021034
4	5	0	1	1	0	0	1	0	1	2.705477	0.7887621	0.7715182	17.776952	-151.66012	-141.71520	3.505131
5	6	1	1	1	0	0	1	0	1	2.082008	0.8374413	0.8205081	5.540639	-163.80517	-151.87127	2.782713
5	6	1	1	1	0	1	0	0	1	2.102923	0.8358083	0.8187050	6.018212	-163.26542	-151.33152	2.738932
5	6	1	1	1	1	0	0	0	1	2.137098	0.8331399	0.8157587	6.798576	-162.39490	-150.46099	2.829352
6	7	1	1	1	0	1	1	0	1	2.005225	0.8434363	0.8234494	5.787389	-163.83429	-149.91140	2.772325
6	7	1	1	1	0	0	1	1	1	2.059621	0.8391892	0.8186601	7.029456	-162.38896	-148.46607	2.839169
6	7	1	1	1	1	0	1	0	1	2.065608	0.8387217	0.8181330	7.166172	-162.23220	-148.30932	2.874944
7	8	1	1	1	0	1	1	1	1	1.972032	0.8460279	0.8225974	7.029455	-162.73565	-146.82378	2.808705
7	8	1	1	1	1	1	1	0	1	2.002941	0.8436146	0.8198168	7.735230	-161.89584	-145.98397	2.882665
7	8	1	1	1	1	0	1	1	1	2.044284	0.8403866	0.8160976	8.679263	-160.79256	-144.88069	2.943495
8	9	1	1	1	1	1	1	1	1	1.970742	0.8461286	0.8187737	9.000000	-160.77098	-142.87013	2.931232

Stepwise regression model (both directions)

- The stepwise regression first fits a SLR for each of the $p-1$ X variables.

$$t_k^* = b_k / s\{b_k\}$$

- The X variable with the largest t^* value is the candidate for the first addition. If the t^* is large, or the P-value is less than some predetermined α, the X variable is added.

- The second variable is added to the model with the first variable. The T test or Partial F test is obtained to determine its significance.

$$F_{new}^* = \frac{MSR(X_{new}|X_{old})}{MSE(X_{old}, X_{new})} \text{ or } t_{new}^* = \sqrt{F_{new}^*}$$

- After two variables are added, the algorithm examines whether any of the variables already in the model should now be dropped.

- Continue till no further variables can either be added or deleted, then the process terminates. AIC is computed for each model in each step

- Since variable can be added and/or removed in each step, this is also the “both” stepwise.
  - If variable can only be added in each step, the method is called “forward” stepwise.
  - If variable can only be removed in each step, the method is called “backward” stepwise.

Stepwise regression model

```
step(lm(lny~blood+prog+enz+liver+age+gender+x7+x8, data=sur), method="both", trace=1)
```

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)

X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

$$\ln(y) \sim blood + prog + enz + age + gender + X8$$

```
Start:  AIC=-160.77
lny ~ blood + prog + enz + liver + age + gender + x7 + x8

          Df Sum of Sq    RSS     AIC
- liver    1   0.00129 1.9720 -162.74
- x7       1   0.03220 2.0029 -161.90
- age      1   0.07354 2.0443 -160.79
<none>                 1.9707 -160.77
- gender   1   0.08415 2.0549 -160.51
- blood    1   0.31809 2.2888 -154.69
- x8       1   0.84573 2.8165 -143.49
- prog     1   2.09045 4.0612 -123.72
- enz      1   2.99085 4.9616 -112.91

Step:  AIC=-162.74
lny ~ blood + prog + enz + age + gender + x7 + x8

          Df Sum of Sq    RSS      AIC
- x7       1    0.0332 2.0052 -163.834
<none>                 1.9720 -162.736
- age      1    0.0876 2.0596 -162.389
- gender   1    0.0971 2.0691 -162.141
- blood    1    0.6267 2.5988 -149.833
- x8       1    0.8446 2.8166 -145.486
- prog     1    2.6731 4.6451 -118.471
- enz      1    5.0986 7.0706  -95.784

Step:  AIC=-163.83
lny ~ blood + prog + enz + age + gender + x8

          Df Sum of Sq    RSS      AIC
<none>                 2.0052 -163.834
- age      1    0.0768 2.0820 -163.805
- gender   1    0.0977 2.1029 -163.265
- blood    1    0.6282 2.6335 -151.117
- x8       1    0.9002 2.9055 -145.809
- prog     1    2.7626 4.7678 -119.064
- enz      1    5.0801 7.0853  -97.672
```

Model Selection Guideline Continued

- It is important to consider the fundamental nature of the explanatory variables.
  - For example, all indicator variables that define a qualitative predictor should be retained in the model.
  - In situations where second order terms X_k^2 or interaction terms $X_k X_m$ are necessary, it is generally recommended to also include the first-order terms, such as X_k and X_m.
- It is crucial to carefully consider the relevance and significance of each variable and to avoid overfitting the model to the data.

For example, after the stepwise algorithm

The algorithm suggests $\ln(y) \sim blood + prog + enz + age + gender + X8$

We should suggest $\ln(y) \sim blood + prog + enz + age + gender + X7 + X8$

	p	1	2	3	4	5	6	7	8	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
1	2	0	0	1	0	0	0	0	0	7.331575	0.4275662	0.4165579	117.409441	-103.8269	-99.84889	8.326716
2	3	0	1	1	0	0	0	0	0	4.312491	0.6632899	0.6500855	50.471575	-130.4833	-124.51634	5.065339
3	4	0	1	1	0	0	0	0	1	2.842883	0.7780337	0.7647157	18.914496	-150.9849	-143.02899	3.469403
4	5	1	1	1	0	0	0	0	1	2.178799	0.8298840	0.8159970	5.750774	-163.3514	-153.40643	2.737771
5	6	1	1	1	0	0	1	0	1	2.082008	0.8374413	0.8205081	5.540639	-163.8052	-151.87127	2.782713
6	7	1	1	1	0	1	1	0	1	2.005225	0.8434363	0.8234494	5.787389	-163.8343	-149.91140	2.772325
7	8	1	1	1	0	1	1	1	1	1.972032	0.8460279	0.8225974	7.029455	-162.7356	-146.82378	2.808705
8	9	1	1	1	1	1	1	1	1	1.970742	0.8461286	0.8187737	9.000000	-160.7710	-142.87013	2.931232

```
library(ALSM)
reducedm<-lm(lny~blood+prog+enz+age+gender+x7+x8, data=sur)
fullm<-lm(lny~blood+prog+enz+liver+age+gender+x7+x8, data=sur)
cpc(reducedm, fullm)
AICp(reducedm)
SBCp(reducedm)
pressc(reducedm)
```

```
[1] 7.029455
[1] -162.7356
[1] -146.8238
[1] 2.808705
```

Model selected by different approaches in the surgical unit example

Y: (ln transformed) survival in days (sur)

X1: blood clotting score (blood)
X2: prognostic index (prog)
X3: enzyme function (enz)
X4: liver function test (liver)
X5: age (age)
X6: gender (gender 0=male, 1=female)
X7: history of alcohol use (X7, 1=moderate, 0=otherwise)
X8: history of alcohol use (X8, 1= severe, 0=otherwise)

	p	1	2	3	4	5	6	7	8	SSEp	r2	r2.adj	Cp	AICp	SBCp	PRESSp
	1	2	0	0	1	0	0	0	0	0	7.331575	0.4275662	0.4165579	117.409441	-103.8269	-99.84889	8.326716
	2	3	0	1	1	0	0	0	0	0	4.312491	0.6632899	0.6500855	50.471575	-130.4833	-124.51634	5.065339
	3	4	0	1	1	0	0	0	0	1	2.842883	0.7780337	0.7647157	18.914496	-150.9849	-143.02899	3.469403
Model 1	4	5	1	1	1	0	0	0	0	1	2.178799	0.8298840	0.8159970	5.750774	-163.3514	-153.40643	2.737771
Model 2	5	6	1	1	1	0	0	1	0	1	2.082008	0.8374413	0.8205081	5.540639	-163.8052	-151.87127	2.782713
Model 3	6	7	1	1	1	0	1	1	0	1	2.005225	0.8434363	0.8234494	5.787389	-163.8343	-149.91140	2.772325
Model 4 (stepwise)	7	8	1	1	1	0	1	1	1	1	1.972032	0.8460279	0.8225974	7.029455	-162.7356	-146.82378	2.808705
	8	9	1	1	1	1	1	1	1	1	1.970742	0.8461286	0.8187737	9.000000	-160.7710	-142.87013	2.931232

X7 can also be added as a option because X7 and X8 both belong to one indicator.

$$Model1: \quad \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_8 X8 \quad + \beta_7 X7$$

$$Model2: \quad \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_6 gender + \beta_8 X8 \quad + \beta_7 X7$$

$$Model3: \quad \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_8 X8 + \beta_7 X7$$

$$Model4: \quad \ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_7 X7 + \beta_8 X8$$

Model Validation: K-fold Cross-validation

- A method for cross-validation that aims to reduce similarity among training datasets is k-fold cross validation. This method involves the following steps:

    1. Randomly split the dataset into k parts, or folds, where k is typically 5 or 10.
    2. Fit the model using k-1 folds, and then calculate the predictive mean squared error (MSE) for the remaining testing set.
    3. Repeat steps 1 and 2 k times, using each fold as the testing set exactly once.
    4. Take the average MSE over the k folds to obtain an estimate of the generalization error of the model.

- K-fold cross-validation is a widely used technique for assessing the performance of a model and selecting optimal hyperparameters. It can help to reduce overfitting by providing a more realistic estimate of the model's performance on unseen data. The choice of the number of folds (k) should depend on the size and complexity of the dataset and the computational resources available.

K-fold Cross-validation result

```r
library(MASS)
library(leaps)
library(caret)

set.seed(123) #set seed for reproducibility

train.control<-trainControl(method="cv", number=10)  #10 fold cross validation

step.model1<-train(lny~blood+prog+enz+x7+x8, data=sur, method="leapBackward",
                   tuneGrid=data.frame(nvmax=5),
                   trControl=train.control)
step.model1$results

step.model2<-train(lny~blood+prog+enz+gender+x7+x8, data=sur, method="leapBackward",
                   tuneGrid=data.frame(nvmax=6),
                   trControl=train.control)
step.model2$results

step.model3<-train(lny~blood+prog+enz+age+gender+x7+x8, data=sur, method="leapBackward",
                   tuneGrid=data.frame(nvmax=7),
                   trControl=train.control)
step.model3$results
```

$Model1$: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_8 X8 + \beta_7 X7$

Root MSE (RMSE)=0.218

$Model2$: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_6 gender + \beta_8 X8 + \beta_7 X7$

Root MSE (RMSE)=0.225

$Model3$: $\ln(y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_5 age + \beta_6 gender + \beta_8 X8$

Root MSE (RMSE)=0.222