Machine learning for data analysis

Week 3 assignment

Running a lasso regression analysis

I chose addhealth as my data set. I calculated the body mass index (BMI) as the quantitative response variable.

The SURVEYSELECT Procedure

As shown in the following table, 70% of the observations were randomly selected as the training set.

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	NEW
Random Number Seed	100
Sampling Rate	0.7
Sample Size	4553
Selection Probability	0.700031
Sampling Weight	0
Output Data Set	TRAINTEST

The GLMSELECT Procedure

The GLMSELECT Procedure

Data Set	WORK.TRAINTEST
Dependent Variable	BMI
Selection Method	LAR
Stop Criterion	None
Choose Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	20
Effect Hierarchy Enforced	None
Random Number Seed	100

Number of Observations Read	6504
Number of Observations Used	6503
Number of Observations Used for Training	4553
Number of Observations Used for Testing	1950

Dimensions	
Number of Effects	8
Number of Parameters	8

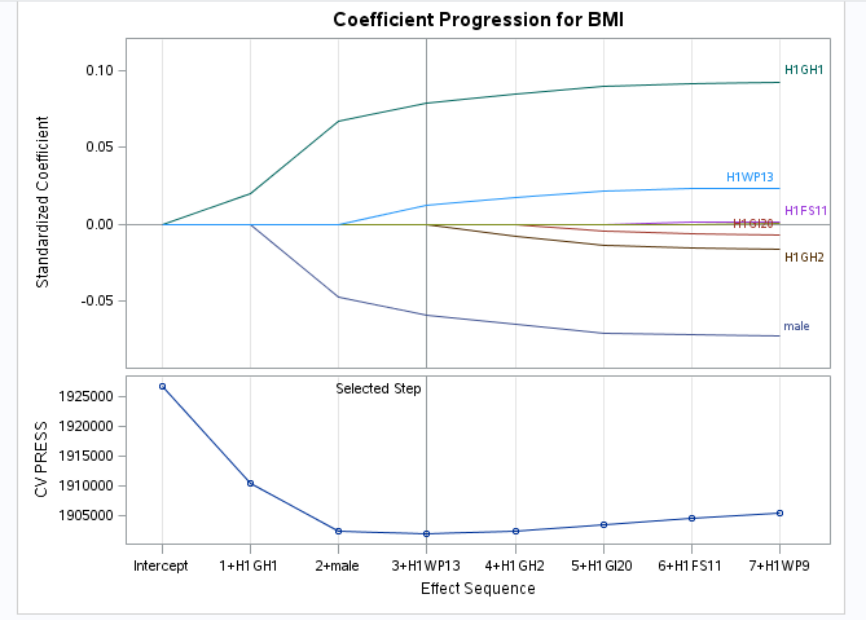
20 folds cross validation was run.

The GLMSELECT Procedure

LAR Selection Summary

Step	Effect Entered	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept	1	422.9660	363.8631	1926844.30
1	H1GH1	2	421.5314	362.4206	1910567.63
2	male	3	417.5405	358.7380	1902382.49
3	H1WP13	4	417.0044	358.4231	1902017.11*
4	H1GH2	5	416.8245	358.3602	1902414.30
5	H1GI20	6	416.7382	358.4270	1903602.94
6	H1FS11	7	416.7298	358.5067	1904601.60
7	H1WP9	8	416.7292	358.5335	1905600.05

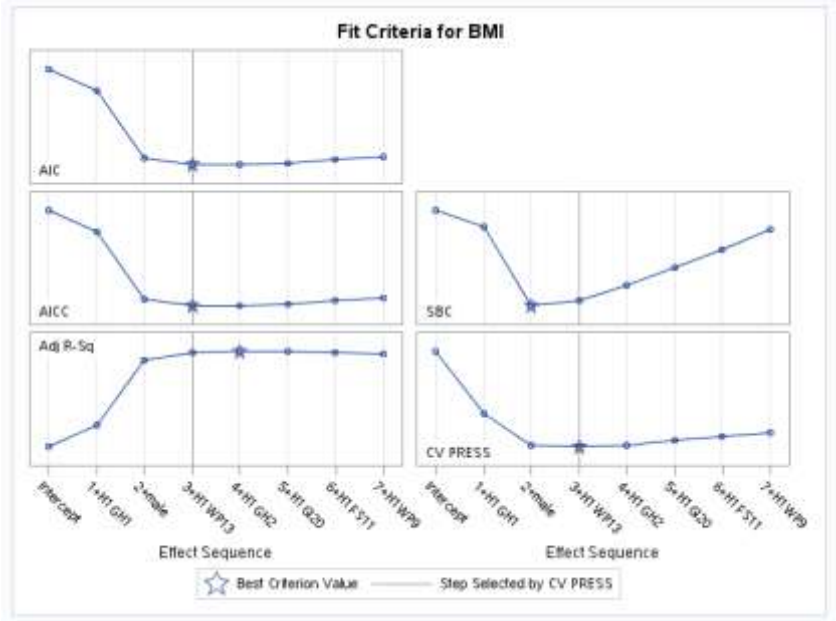
* Optimal Value of Criterion



As more variables are added into the model, the ASE and Test ASE decrease. The best model appears when H1WP13 is added into the model.

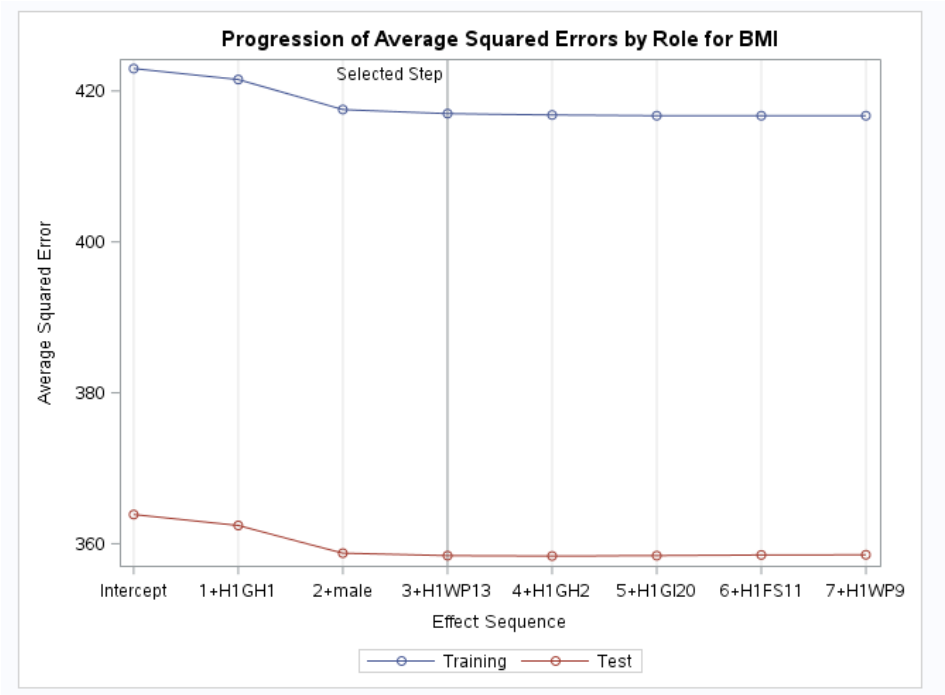
The first most important variables are H1GH1, male and H1WP13 respectively.

H1GH1 and H1WP13 are positively correlated to BMI, while male and H1GH2 are negatively correlated to BMI.



Different criteria including AIC, AICC and CV PRESS choose H1GH1, male and H1WP13 as the best model.

SBC chooses H1GH1 and male as the best model. While Adj R-sq chooses the H1GH1, male, H1WP13 and H1GH2 as the best model.



The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 3.

Effects:	Intercept male H1GH1 H1WP13

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	27143	9047.73145	21.68
Error	4549	1898621	417.37106	
Corrected Total	4552	1925764		

Root MSE	20.42966
Dependent Mean	24.86094
R-Square	0.0141
Adj R-Sq	0.0134
AIC	32032
AICC	32032
SBC	27502
ASE (Train)	417.00438
ASE (Test)	358.42308
CV PRESS	1902017

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	21.453433

This shows the best model.

Code:

```
*Load data;
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;

H1GH59=H1GH59A * 12 + H1GH59B;/*add a new variable*/
BMI=H1GH60 * 0.454/(H1GH59 * 0.0254)**2;/*body mass index*/


label AID="respondent ID"
      BIO_SEX="gender"
      H1GI20="grade"
      H1GH1="general health"
      H1GH2="frequency of headache"
      H1GH6="frequnecy of feeling weak"
      H1GH59A="height in feet"
      H1GH59B="height in inch"
      H1GH60="weigt (pound)"
      H1FS11="feeling happy"
      H1WP9="how close with mother"
      H1WP13="how close with father"
      H1GH59="height (inch)";

if BIO_SEX = 1 then male = 1;
if BIO_SEX = 2 then male = 0;

keep AID male BMI H1GI20 H1GH1 H1GH2 H1FS11 H1WP9 H1WP13;

```

```
proc sort; by AID;


ods graphics on;

*split data;
proc surveyselect data=new out=traintest seed=100
                  samprate = 0.7 method = srs outall;
run;

*lasso regression;
proc glmselect data= traintest plots=all seed=100;
  partition role=selected (train = '1' test = '0');
  model BMI = male H1GI20 H1GH1 H1GH2 H1FS11 H1WP9 H1WP13/selection=lar (choose = cv stop = none)
                                                          cvmethod=random(20);
run;
```