Model Building with Variable Selection – Part IV

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 10 – Lecture 10B

Stepwise Regression Models: Forward Selection

Start with no regression.

Enter one regressor with largest simple correlation with y or the largest value of F statistic for testing significance of regression, (say F-to-enter, labeled F_{IN}).

Enter the next regressor with the largest

$$F = \frac{SS_R(x_2 \mid x_1)}{MS_{\text{Res}}(x_1, x_2)} > F_{\text{IN}}$$

Continue this process

The Hald Cement Data: Forward Selection

Stepwise Regression: y versus x1, x2, x3, x4

Forward selection. Alpha-to-enter: 0.25

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	117.57	103.10	71.65
x4	-0.738	-0.614	-0.237
T-Value	-4.77	-12.62	-1.37
P-Value	0.001	0.000	0.205
x1		1.44	1.45
T-value		10.40	12.41
P-Value		0.000	0.000
x2			0.42
T-Value			2.24
R-Value			0.052
S	8.96	2.73	2.31
R-Sq	67.45	97.25	98.23
R-Sq(adj)	64.50	96.70	97.64
Mallows C-p	138.7	5.5	3.0

< -1.21

>1.22

>1.23

Figure 10.8 Forward selection results from Minitab for the Hald cement data.

$t_{.25,11}=1.21$ $t_{.25,10}=1.22$ $t_{.25,9}=1.23$

Stepwise Regression Models: Backward Elimination

Start with full regression.

Remove the regressor with smallest value of F statistic less than F_{OUT}.

Continue this process until the smallest F value is not less than F_{OUT}.

The Hald Cement Data: Backward Selection

Stepwise Regression: y versus x1, x2, x3, x4

Backward elimination. Alpha-to-Remove: 0.1

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	62.41	71.65	52.58
x1	1.55	1.45	1.47
T-Value	2.08	12.41	12.10
P-Value	0.071	0.000	0.000
x2	0.510	0.416	0.662
T-Value	0.70	2.24	14.44
P-Value	0.501	0.052	0.000
x3	0.10		
T-Value	0.14		
P-Value	0.896		
x4	-0.14	-0.24	
T-Value	-0.20	-1.37	
P-Value	0.844	0.205	
S	2.45	2.31	2.41
R-Sq	98.24	98.23	97.87
R-Sq(adj)	97.36	97.64	97.44
Mallows C-p	5.0	3.0	2.7

<1.86

> -1.83

Figure 10.9 Backward selection results from Minitab for the Hald cement data.

$t_{.10,8}=1.86$ $t_{.10,9}=1.83$ $t_{.10,10}=1.81$

Stepwise Regression

Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their F (or t) statistics.

A regressor added at an earlier step may now be redundant because of the relationships between it and regressors now in the equation. If the F (or t) statistic for a variable is less than F_{OUT} (or t_{OUT}), the variable is dropped from the model.

Stepwise Regression

Stepwise regression requires two cutoff values: F_{IN} and F_{OUT}; often $F_{IN} > F_{OUT}$

Strategy for variable selection and model building (Figure 10.11 – Flowchart, p.351)

Case 10.4 (p.354-366)

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING