Automatic Variable Selection

```
title2 'Automatic variable selection: Goal is to predict grade';
%include 'readmath.sas';
options pagesize=1000; /* Print fewer page headings */

/* The data step continues */
if ethnic ne 6;
if course ne 4;   /* Otherwise, throw the case out */

/* Dummy variables for ethnic background */
if ethnic=. then e1=.;
   else if ethnic=1 then e1=1;
   else e1=0;
if ethnic=. then e2=.;
   else if ethnic=2 then e2=1;
   else e2=0;
if ethnic=. then e3=.;
   else if ethnic=3 then e3=1;
   else e3=0;
if ethnic=. then e4=.;
   else if ethnic=4 then e4=1;
   else e4=0;

/* Ethnic
                1 = 'Asian'
                2 = 'Eastern European'
                3 = 'European not Eastern'
                4 = 'Middle-Eastern  and Pakistani'
                5 = 'East Indian'
                6 was deleted.
*/
label e1 = 'Asian vs East Ind.'
      e2 = 'East Eur. vs East Ind.'
      e3 = 'Other Eur. vs East Ind.'
      e4 = 'Mid. East & Pak. vs East Ind.';

if sex = 'Female' then gender=1; else if sex = 'Male' then gender=0;
if tongue = 'English' then mtongue=1; else if tongue='Other' then mtongue=0;

if course=. then c1=.; else if course=1 then c1=1; else c1=0;
if course=. then c2=.; else if course=2 then c2=1; else c2=0;

proc freq;
     title3 'Check dummy variables for course';
     tables (c1 c2) * course / norow nocol nopercent missing;

proc stepwise;
     title3 'Forward selection';
     model grade = gender mtongue e1-e4
                   hsgpa hscalc hsengl
                   c1 c2 precalc calc totscore
                   / forward slentry=0.05;
                          /* Default signif level for entry is 0.5 */
```

```
proc stepwise;
     title3 'Stepwise (mixed) selection';
     model grade = gender mtongue e1-e4
                   hsgpa hscalc hsengl
                   c1 c2 precalc calc totscore
                   / stepwise slentry=0.05 slstay=0.05;
                          /* Default slentry = slstay = 0.15 */

proc reg;
     title3 'Test English language variables together';
     model grade = hsgpa hscalc hsengl  totscore  mtongue;
     English: test hsengl=mtongue=0;

proc iml;
     title3 'Proportion of remaining varation';
     print "hsengl and mtongue controlling for hsgpa, hscalc, totscore";
     F = 6.80; s = 2; dfe = 274;      /* dfe = n-p */
     a = s*F / (dfe + s*F); print a;
```

`Proc freq` output, before adding `if course ne 4;`

Table of c1 by course

c1 / course	.	1	2	3	No Resp	Total
.	79	0	0	0	0	79
0	0	0	363	39	4	406
1	0	54	0	0	0	54
Total	79	54	363	39	4	539

Table of c2 by course

c2 / course	.	1	2	3	No Resp	Total
.	79	0	0	0	0	79
0	0	54	0	39	4	97
1	0	0	363	0	0	363
Total	79	54	363	39	4	539

Forward selection

The STEPWISE Procedure
Model: MODEL1
Dependent Variable: grade Final mark (if any)

Number of Observations Read	535
Number of Observations Used	280
Number of Observations with Missing Values	255

Forward Selection: Step 1

Variable hsgpa Entered: R-Square = 0.3537 and C(p) = 57.2365

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	33894	33894	152.17	<.0001
Error	278	61921	222.73759		
Corrected Total	279	95815			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-87.40438	12.04376	11731	52.67	<.0001
hsgpa	1.82819	0.14820	33894	152.17	<.0001

Bounds on condition number: 1, 1

Values of *C(p)* are supposed to be better when they are small, and when they are close to *p* -- the number of regression coefficients.

Forward Selection: Step 2

Variable totscore Entered: R-Square = 0.4120 and C(p) = 29.2151

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	39473	19736	97.03	<.0001
Error	277	56343	203.40270		
Corrected Total	279	95815			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-77.36795	11.66763	8943.62142	43.97	<.0001
hsgpa	1.55985	0.15061	21819	107.27	<.0001
totscore	1.32372	0.25276	5578.50005	27.43	<.0001

Bounds on condition number: 1.1309, 4.5236

Forward Selection: Step 3

Variable hsengl Entered: R-Square = 0.4414 and C(p) = 16.0143

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	42297	14099	72.71	<.0001
Error	276	53518	193.90578		
Corrected Total	279	95815			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-72.63764	11.45922	7791.20117	40.18	<.0001
hsgpa	1.94884	0.17892	23006	118.65	<.0001
hsengl	-0.45289	0.11866	2824.55260	14.57	0.0002
totscore	1.17049	0.25004	4249.35806	21.91	<.0001

Bounds on condition number: 1.6742, 12.964

Forward Selection: Step 4

Variable mtongue Entered: R-Square = 0.4530 and C(p) = 12.0592

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	43404	10851	56.93	<.0001
Error	275	52411	190.58700		
Corrected Total	279	95815			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-68.15382	11.51212	6679.79905	35.05	<.0001
mtongue	-5.19142	2.15448	1106.57185	5.81	0.0166
hsgpa	1.88196	0.17954	20941	109.88	<.0001
hsengl	-0.38223	0.12124	1894.23591	9.94	0.0018
totscore	1.13240	0.24839	3961.19628	20.78	<.0001

Bounds on condition number: 1.7152, 22.144

Forward Selection: Step 5

Variable hscalc Entered: R-Square = 0.4616 and C(p) = 9.6135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44230	8846.02865	46.99	<.0001
Error	274	51585	188.26767		
Corrected Total	279	95815			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-66.74563	11.46159	6384.57537	33.91	<.0001
mtongue	-4.71674	2.15329	903.34525	4.80	0.0293
hsgpa	1.59500	0.22496	9464.04494	50.27	<.0001
hscalc	0.21769	0.10392	826.08202	4.39	0.0371
hsengl	-0.30921	0.12544	1143.87785	6.08	0.0143
totscore	0.98959	0.25612	2810.72834	14.93	0.0001

Bounds on condition number: 2.726, 44.206

No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square
1	hsgpa	High School GPA	1	0.3537	0.3537
2	totscore	Total # right on diagnostic test	2	0.0582	0.4120
3	hsengl	HS English	3	0.0295	0.4414
4	mtongue		4	0.0115	0.4530
5	hscalc	HS Calculus	5	0.0086	0.4616

Summary of Forward Selection

Step	C(p)	F Value	Pr > F
1	57.2365	152.17	<.0001
2	29.2151	27.43	<.0001
3	16.0143	14.57	0.0002
4	12.0592	5.81	0.0166
5	9.6135	4.39	0.0371

Mixed forward and backwards selection gave exactly the same results we've just seen, because no variables were removed. Output has been deleted.

Test English language variables together

The REG Procedure
Model: MODEL1
Dependent Variable: grade Final mark (if any)

Number of Observations Read	535
Number of Observations Used	280
Number of Observations with Missing Values	255

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44230	8846.02865	46.99	<.0001
Error	274	51585	188.26767		
Corrected Total	279	95815			

Root MSE	13.72107	R-Square	0.4616
Dependent Mean	60.75714	Adj R-Sq	0.4518
Coeff Var	22.58346		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	-66.74563	11.46159
hsgpa	High School GPA	1	1.59500	0.22496
hscalc	HS Calculus	1	0.21769	0.10392
hsengl	HS English	1	-0.30921	0.12544
totscore	Total # right on diagnostic test	1	0.98959	0.25612
mtongue		1	-4.71674	2.15329

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	-5.82	<.0001
hsgpa	High School GPA	1	7.09	<.0001
hscalc	HS Calculus	1	2.09	0.0371
hsengl	HS English	1	-2.46	0.0143
totscore	Total # right on diagnostic test	1	3.86	0.0001
mtongue		1	-2.19	0.0293

Test English language variables together

The REG Procedure
Model: MODEL1

Test English Results for Dependent Variable grade

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1280.91501	6.80	0.0013
Denominator	274	188.26767		

Proportion of remaining varation

hsengl and mtongue controlling for hsgpa, hscalc, totscore

a

0.0472879