Model Selection: Backward & Stepwise Procedures—Water Level Study

A. Introduction.

Observations on nine explanatory (independent) variables were obtained in the 'water Level Study'**:** sex gravity totphys bryant vander triangle trailer tree comphys, and two new variables were created from these: (i) moving and (ii) total. The dependent variable was y: (pass or fail the water level task). Two goals of the study were:

1 Which subset of the variables are statistically significantly related to passing/failing the water level task?
2. Can the difference between females and males on the water level task be explained by the independent variables?

In the following we look at these two issues. We begin by using two subset selection procedures in SAS Proc Logistic for choosing variables related to the response:

1. Backward elimination
2. Stepwise selection

B. SAS Program:

```
options ls=72;
data water;
input obs y sex gravity totphys bryant vander triangle trailer tree
comphys moving total;
cards;
   1    0    1    4    5    3    10    0    6    1    1    1    25
   2    1    2    5    9    6    12    0    6    4    4    1    37
 ...    .    .    .    .    .    ..    .    .    .    .    .    ..
 ...    .    .    .    .    .    ..    .    .    .    .    .    ..
 166    0    1    4    7    5    12    2    6    3    3    1    35
;

Proc Logistic; Model Y=sex gravity totphys bryant vander triangle
trailer tree comphys moving/backward;
Proc Logistic; Model Y=sex gravity totphys bryant vander triangle
trailer tree comphys moving/stepwise;
run;
```

C. Model Selection: Backward Elimination

The procedure goes in steps;

Step 0. Fit the model with all 10 variables included. The value of -2ln L (-2 ln likelihood) = 123.105. The LRT value for testing all parameters are simultaneously 0 is LRT = 236.036-123.104 = 102.932.

Step 1. In Step 0, identify the variable with the smallest G^2 for testing its' parameter is 0, adjusted for all other variables in the model—this is (always?) the value with smallest Wald Chi-square or smallest p-value. In this case, the variable is 'triangle'. Calculate the

change in deviance with the value in and out of the model: $G^2 = 102.932-102.897 = 0.035$.
The criterion for significance is $G^2 > 3.84 = \chi^2(1, .05)$. We conclude that 'triangle' is not a significant variable, adjusted for the other variables in the model.

Step 2. Omit the variable identified as not being 'significant' in Step 1. Re-run the logistic regression model with 'triangle' deleted. Identify the variable with the smallest G^2 for testing its' parameter is 0, adjusted for all other variables in the model, as in Step 1. In this case, the variable is 'totphys'. Calculate the change in deviance with the value in and out of the model: $G^2 = 102.897-101.660 = 1.237$.
The criterion for significance is $G^2 > 3.84 = \chi^2(1, .05) = 3.84$. We conclude that 'totphys' is not a significant variable, adjusted for the other variables in the model

Continue until no variable, adjusted for others in the model, meets the criterion for deletion.

SAS OUTPUT: Backward Elimination:

The LOGISTIC Procedure

Model Information

Data Set	WORK.WATER
Response Variable	y
Number of Response Levels	2
Number of Observations	166
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	y	Total Frequency
1	0	96
2	1	70

Probability modeled is y=0.

Stepwise Selection Procedure

Step 0. The following effects were entered:

Intercept sex gravity totphys bryant vander triangle trailer tree comphys moving

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	145.104
SC	231.148	179.336
-2 Log L	226.036	123.104

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq

Likelihood Ratio	102.9319	10	<.0001
Score	75.9882	10	<.0001
Wald	39.8693	10	<.0001

Step 1. Effect triangle is removed:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	143.139
SC	231.148	174.259
-2 Log L	226.036	123.139

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	102.8970	9	<.0001
Score	75.3948	9	<.0001
Wald	39.8243	9	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.0345	1	0.8527

Step 2. Effect totphys is removed:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	142.376
SC	231.148	170.384
-2 Log L	226.036	124.376

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	101.6598	8	<.0001
Score	74.7438	8	<.0001
Wald	39.6571	8	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.9347	2	0.6267

Step 3. Effect comphys is removed:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	141.542
SC	231.148	166.438
-2 Log L	226.036	125.542

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	100.4940	7	<.0001
Score	73.5038	7	<.0001
Wald	37.9008	7	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
2.0543	3	0.5612

Step 4. Effect sex is removed:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	142.385
SC	231.148	164.169
-2 Log L	226.036	128.385

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	97.6504	6	<.0001
Score	70.8081	6	<.0001
Wald	37.6798	6	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4.8313	4	0.3050

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Removed	Number DF	In	Wald Chi-Square	Pr > ChiSq
1	triangle	1	9	.0345	0.8527
2	totphys	1	8	0.8349	0.3609
3	comphys	1	7	1.1476	0.2840
4	sex	1	6	2.8044	0.0940

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	7.8773	2.2166	12.6295	0.0004
gravity	1	-0.5583	0.1783	9.8099	0.0017
bryant	1	-0.3691	0.1737	4.5149	0.0336
vander	1	-0.2044	0.0712	8.2388	0.0041
trailer	1	-0.7125	0.2961	5.7890	0.0161
tree	1	-0.4932	0.1642	9.0239	0.0027
moving	1	2.4148	0.9041	7.1340	0.0076

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
gravity	0.572	0.403	0.811
bryant	0.691	0.492	0.972
vander	0.815	0.709	0.937
trailer	0.490	0.274	0.876
tree	0.611	0.443	0.842
moving	11.187	1.902	65.808

C2. Model Selection: Stepwise Selection.

Step 0. Fit the intercept model (this step is really not needed). -2ln L = 226.036

Step 1. Fit all models (10 in all) with each variable in the model. Identify that model which has the highest -2lnL. In this case, the model with 'totphys' in it has -2lnL = 178.992—all other models with one variable in them have -2lnL < 178.992. The test of significance of the variable is given by $G^2 = 226.036\text{-}178.992 = 47.044$. We reject the hypothesis that the parameter associated with 'totphys' is 0 if $G^2 > 3.84 = \chi^2(1, .05)$. We conclude that 'totphys' is a statistically significant variable (predictor)

Step 2. Run all 2 variable models with 'totphys' and one other predictor. Identify that variable which, together with 'totphys', has the highest value of -2lnL (or equivalently, the highest value of G^2 for testing the global hypothesis that the two predictors are not jointly significant. In this example, the variable is 'vander'. The value of -2lnL is 159.484. Then test for significance of the addition of 'vander' to the model containing 'totphys'. The change in -2lnL, or G^2, is given by -2lnL(totphys) – [-2lnL(totphys, vander)] = 178.992-159.484 = 19.508.

Step 3. Proceed as in steps 1 and 2, to identify the third variable, which together with totphys and vander, yields the highest -2lnL or the biggest change in G^2. This variable is 'tree', with -2lnL = 148.637. The change in deviance is 159.484-148.637 = 10.847. The change is statistically significant.

Now check to see if one of the three variables totphys, vander, or tree, can be deleted from the model without a significant decrease in deviance. This variable could only be 'totphys' (for logical reasons). It turns out that it cannot be deleted.

Continue to add variables and checking to see if any can be deleted, until none can be added. Stop.

SAS OUTPUT: Stepwise Selection Procedure

Step 0. Intercept entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
75.9882	10	<.0001

Step 1. Effect totphys entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	182.993
SC	231.148	189.217

-2 Log L	226.036	178.993

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	47.0426	1	<.0001
Score	41.5656	1	<.0001
Wald	32.9603	1	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
44.3755	9	<.0001

Step 2. Effect vander entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	165.484
SC	231.148	174.820
-2 Log L	226.036	159.484

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	66.5513	2	<.0001
Score	57.2050	2	<.0001
Wald	41.6365	2	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
28.3054	8	0.0004

Step 3. Effect tree entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	156.637
SC	231.148	169.085
-2 Log L	226.036	148.637

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	77.3983	3	<.0001
Score	64.0331	3	<.0001
Wald	42.4435	3	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
20.8917	7	0.0039

Step 4. Effect trailer entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	150.284
SC	231.148	165.844
-2 Log L	226.036	140.284

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	85.7518	4	<.0001
Score	66.7439	4	<.0001
Wald	40.2427	4	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
15.3522	6	0.0177

Step 5. Effect moving entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	143.792
SC	231.148	162.463
-2 Log L	226.036	131.792

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	94.2442	5	<.0001
Score	69.0294	5	<.0001
Wald	39.2090	5	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
8.0348	5	0.1543

Step 6. Effect bryant entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	228.036	141.725
SC	231.148	163.509
-2 Log L	226.036	127.725

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	98.3107	6	<.0001
Score	71.9151	6	<.0001
Wald	40.1958	6	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4.0274	4	0.4023

NOTE: No (additional) effects met the 0.05 significance level for entry

into the model.

Summary of Stepwise Selection

Step	Effect Entered	Removed	Number DF	In	Score Chi-Square	Pr > ChiSq
1	totphys		1	1	41.5656	<.0001
2	vander		1	2	19.1098	<.0001.
3	tree		1	3	10.6764	0.0011
4	trailer		1	4	6.6594	0.0099
5	moving		1	5	7.5976	0.0058
6	bryant		1	6	3.9691	0.0463

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard Estimate	Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	7.7840	2.1401	13.2297	0.0003
totphys	1	-0.3914	0.1209	10.4818	0.0012
bryant	1	-0.3412	0.1739	3.8479	0.0498
vander	1	-0.2059	0.0719	8.2126	0.0042
trailer	1	-0.6802	0.2846	5.7122	0.0168
tree	1	-0.4534	0.1640	7.6388	0.0057
moving	1	2.3768	0.8904	7.1250	0.0076

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
totphys	0.676	0.533	0.857
bryant	0.711	0.506	1.000
vander	0.814	0.707	0.937
trailer	0.507	0.290	0.885
tree	0.635	0.461	0.876
moving	10.771	1.881	61.684

D. Conclusions.

1. Variables Selected and Estimates

Odds Ratio Estimates

Backward elimination

Effect	Point Estimate	95% Wald Confidence Limits	
gravity	0.572	0.403	0.811
bryant	0.691	0.492	0.972
vander	0.815	0.709	0.937
trailer	0.490	0.274	0.876
tree	0.611	0.443	0.842
moving	11.187	1.902	65.808

Stepwise Selection

Effect	Point Estimate	95% Wald Confidence Limits	
totphys	0.676	0.533	0.857
bryant	0.711	0.506	1.000
vander	0.814	0.707	0.937
trailer	0.507	0.290	0.885
tree	0.635	0.461	0.876
moving	10.771	1.881	61.684

The two procedures each selected 6 variables with 5 in common; backward elimination chose ‘gravity’ while stepwise chose ‘totphysics’. The odd ratio and confidence interval estimates are quite close for all variables.

2. Neither model includes ‘sex’. We conclude that adjusted for these 6 independent variables ‘sex’ does not affect passing/failing.